Improving Random Forest Method to Detect Hatespeech and Offensive Word

Kristiawan Nugroho
Doctoral Program
Dian Nuswantoro University
Semarang,Indonesia
kristiawan@mhs.dinus.ac.id

Edy Noersasongko
Faculty of Computer Science
Dian Nuswantoro University
Semarang,Indonesia
edi-nur@dosen.dinus.ac.id

Purwanto
Faculty of Computer Science
Dian Nuswantoro University
Semarang,Indonesia
purwanto@dsn.dinus.ac.id

Muljono
Faculty of Computer Science
Dian Nuswantoro University
Semarang,Indonesia
muljono@dsn.dinus.ac.id

Ahmad Zainul Fanani
Faculty of Computer Science
Dian Nuswantoro University
Semarang,Indonesia
a.zainul.fanani@dsn.dinus.ac.id

Affandy
Faculty of Computer Science
Dian Nuswantoro University
Semarang,Indonesia
affandy@dsn.dinus.ac.id

Ruri Suko Basuki
Faculty of Computer Science
Dian Nuswantoro University
Semarang,Indonesia
ruri.basuki@dsn.dinus.ac.id

Abstract—**Hate Speech is a problem that often occurs when someone communicates with each other using social media on the Internet. Research on hate speech is generally done by exploring datasets in the form of text comments on social media such as Twitter, Facebook and MySpace. This study aims to improve the performance of the Random Forest method in detecting hatespeech and crude speech. In this paper the researcher uses a twitter hate speech and offensive identification dataset that is classified using the Random Forest method which will be compared with the results of its accuracy with AdaBoost and Neural Network to detect hatespeech and crude speech. The detection results of hatespeech and crude speech identification resulted in an accuracy of 0.722 for the Random Forest method and 0.708 using AdaBoost and 0.596 using Neural Network method.**

Keywords-Hate,Speech,Social,Media,Classification,Random Forest,AdaBoost,Neural Network.

I. INTRODUCTION

The development of information technology today, especially the Internet, has brought people in the stages of modern life. The internet makes it easy for humans to communicate with each other both via email, website and social media. However, the development of information technology has had a negative effect with the presence of hate speeches. Hate speeches can be done using electronic media through social media such as Twitter, Facebook or MySpace. Hate speech is a form of communication that contains someone's disparagement or groups based on race, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics [1]. According to data obtained from the legal statistics bureau under the United States Department of Law US Population experienced an average of 250,000 victimization racial crimes each year from 2004 to 2015. Another study supported by UNESCO discussed the problem of increasing hate speech with the emergence of the internet from a legal and social perspective, They said that platforms such as Facebook and Twitter only adopted a reactive approach to handling hate speech reported by users, they should be able to do better [2]. The issue of hate speech in Indonesia, as conveyed by the Indonesian Telematics Society (Mastel), as many as 91.8 percent of respondents claimed to receive the most social and political hoax content, such as regional elections and government, not much different from social politics, the issue of SARA was in a position the second with 88.6%. Hatespeech causes a variety of problems that threaten harmony between humans because of the freedom of speech but in reality it is used to do bullying and utterances of hatred towards others. By looking at the data and facts presented above, we can see that hate speech is a serious problem that must be resolved wisely.

Research on hate speech is still developing until now, as research conducted by Ricardo Martins [3] by classifying hate speech datasets by using Natural Language Processing (NLP). In addition, research on the detection of hate speech through social media uses Random Forest and Naïve Bayes was also carried out by Zewdie and Jenq-Haur Wang [4] with the proposed method producing an accuracy of 79.83%. Research on hate speech in Indonesia began with initial studies and building hate speech datasets using the Bayesian Logistic Regression and random Forest Decision Tree [5]. This paper discusses the detection of hate speech and offensive words

978-1-7281-1655-6/19/$31.00 ©2019 IEEE

using Random Forest, AdaBoost and Neural Network using datasets taken from a collection of comments on Twitter.

II. RELATED WORK

Initially in 2004 researchers only discussed the identification of web pages containing hatred, racism and extremism [6], but nowadays the problem of hate speech is actually more and more happening as the development of social media [7]. The problems regarding the prediction and classification of hate speeches through social media twitter and Facebook are interesting topics that are being researched up to now. Research on hate speech conducted by Kwok and Wang in 2013 focused more on racism on skin color, because the presence of hate speech and harsh words that were relatively high in social media made research on hate speech more challenging [8]. Research on hate speech generally uses a classification approach in machine learning using various methods such as SVM, Decession Tree, Random Forest, Artificial Neural Network and Naive Bayes.

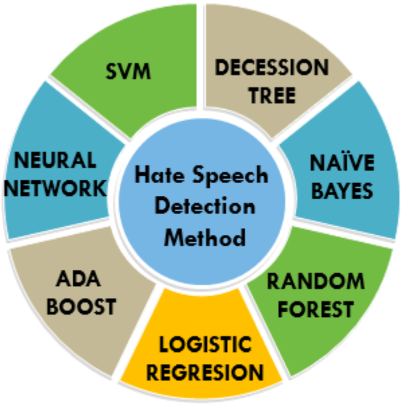


Figure 1 : Hatespeech detection methods

Research on hate speech was carried out by Hema Krishnan regarding speech detection on Twitter using Naïve Bayes Classifier [9], emotional data such as fear, joy, sadness, anger and so on extracted to be included in the MongoDB database. Naufal Riza conducted research on Naïve implementation Bayes in classifying hate speech using Indonesian Language [10]. The Naïve Bayes model is also used by Kelvin Kiema and George Okeyo in conducting hate speech classifications on twitter social media [11], this research produces better performance by producing precision, recall and accuracy respectively at 58%, 62% and 67.47%. In addition to using Naïve Bayes, some researchers also use other methods such as Neural Networks such as those conducted by Venkateswarlu in classifying sound using Neural Network [12]. This research uses Recurent Neural Network and produces a suitable model in classifying sound signals. Besides ANN another method that has proven accuracy in detecting sentiment analysis is the Random Forest method, this method is used by Ali Fauzy [13] in detecting sentiment analysis in Indonesian, Producing 0829 OOB values (Out Of Bag) which indicates that this model has good performance. The use of the Random Forest in solving word prediction problems was also conducted by Sanjana Sharma [14] who predicted dangerous words using Random Forest, the result was 76.42% accuracy in this study, this achievement was better than the 2 models previously used was 72.42% using Naive Bayes and 71.71% using SVM.

Davidson [15] uses Porter stemmer and Logistic Regression, Naïve Bayes, Decession Trees, Random Forest and Linear SVM models, Davidson [15] has produced Linear SVM and Logistic Regression as models that produce levels good accuracy after evaluation uses 5 fold cross validation. Servin Malmasi [16] also conducted research using the same dataset as Davidson used. This study uses the SVM Linear method and testing uses a 10 fold cross. Other research on human emotion detection done by Jasdeep Singh [17] uses AdaBoost and Neural Networks that are proven to be able to detect emotional levels in humans well. This research detects hate speech and offensive words by using Random Forest ,Adaptive Booster (AdaBoost) and Neural Network which will be compared with the results of measurement accuracy so that models can be found that are suitable in recognizing hate speech and harsh words that are often on social media. With the introduction of Hatespeech method which has a high level of accuracy, it is expected to be able to be used in detecting and eliminating comments included in the Hatespeech group on social media so that the problem with many comments in a number of media, especially social media can be reduced and even eliminated to maintain a healthier social media life and safe without hatespeech.

III. METHOD

The method that will be used in conducting detection of hate speech and offensive words in this paper is Random Forest, Adaptive Booster and Neural Network. The detection process can be seen in Figure 2 as follows:

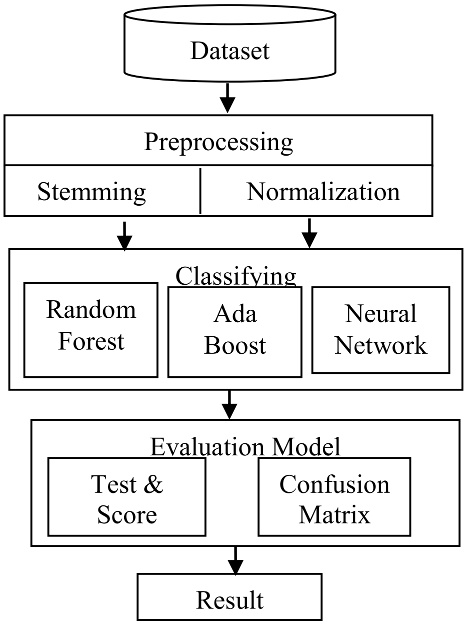


Figure 2 : Hatespeech's Prediction Model

A. *Random Forest*

Random Forest is one of the methods in Machine Learning that is used in the process of classifying large amounts of data. The first Random Forest was introduced by Ho in 1995, Random Forest works by combining many trees in training data so that it will produce a high level of accuracy [18]. Random Forest is a development of the Classification and Regression Tree (CART) method by applying bootstrap aggregating (bagging) and random feature selection methods [19].

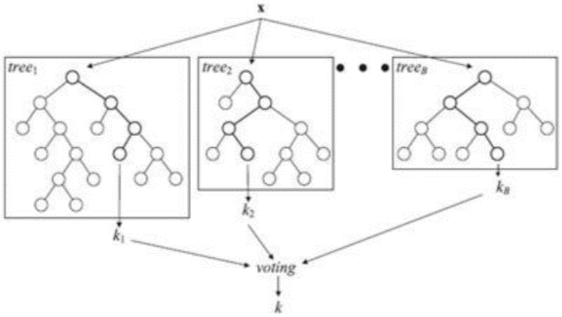


Figure 3 :Random Forest Architecture (Verikas et al.2011)

Random Forest method algorithm according to Breiman and Cutler [20] is as follows:

1. Take a random sample of size n with recovery in the data cluster. This step is called bootstrap (bag).
2. Using the bootstrap example, the tree is built until it reaches the maximum size without pruning. Tree construction is done by applying random feature selection, namely m explanatory variable randomly selected where m << p, then the best sorter is chosen based on $ m $ explanatory variable.
3. Repeat steps 1 and 2 times to create a forest consisting of k trees.In determining the classification in Random Forest is taken based on votes from each tree, the most votes will be the winner. According to Yin, the Random Forest formation uses the Gini Index value to determine the split that will be used as a node [20] with the following formula:

$$\text{Gini (S)}=1 - \sum_{i=1}^{k} pi^2$$

pi is the probability of S belonging to class i

After calculating the Gini value the next step is to calculate the Gini Gain value using the formula:

$$\text{GiniGain (S)= Gini (S) – Gini (A,S) = Gini (S) - }\sum_{i=1}^{n}\frac{|S_i|}{|S|}\text{ Gini }(S_i)$$

where S_i is the partition of S caused by attribute A.

B. *Adaptive Booster (AdaBoost)*

Adaptive Booster is one of the algorithms in machine learning developed by Freund and Schapire [21]. The Adaboost method gives more weight to weak classification. The algorithm of this method can be explained in steps:

1. Minimize the error function with the formula :

$$W_e = \sum_{y_i \neq k_m(x_i)} w_i^{(m)} \exp(\alpha_m)$$

2. Set the value α with the formula:

$$\alpha_m = \frac{1}{2}\ln(\frac{1-e_m}{e_m})\text{ yang mana } e_m = \frac{W_e}{W}$$

3. Update values if observing misclassification by formula:

$$w_i^{(m+1)} = w_i^{(m)} \exp(\alpha_m) = w_i^{(m)}\sqrt{\frac{1-e_m}{e_m}}$$

4. For other values using the formula:

$$w_i^{(m+1)} = w_i^{(m)} \exp(-\alpha_m) = w_i^{(m)}\sqrt{\frac{e_m}{1-e_m}}$$

C. *Neural Network*

Neural Network (NN) is a field of soft computing that studies the mechanism of methods that resemble the capabilities of the human brain that can provide stimulation, process and provide output. One method that is often used in NN is the backpropagation shown in the figure as follows:

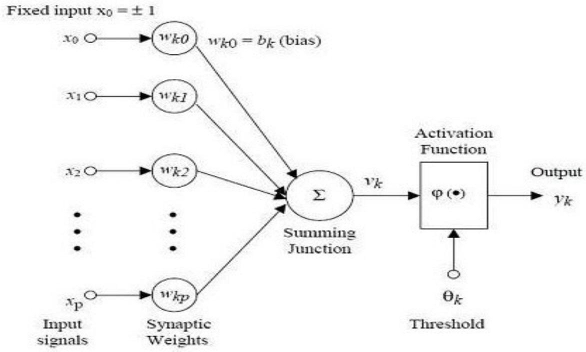


Figure 4 : Neural Network Architecture

In the feed forward process, it is done by calculating the Weighted Sum and sending the amount to an Activation Function to produce Output. The formulas used to calculate weighted sum are:

A_j(X,W) = SUM(i=1 to n) x_i * w_ji

where:
j : index of output
i : index of input
x_i : input to i
w_ji weight from input to i to output to j

For the Activation Function, there are many that can be used, but the most widely used is the Sigmoid Function which is formulated as follows:

O_j(X,W) = 1 / (1 + exp(A_j(X,W)))

where :
O_j(X,W) : output to j

IV. RESULT AND DISCUSSION

A. Dataset

This paper used 14,509 datasets of Hate Speech Identification taken from data.world. This dataset consists of 16 attributes taken by tweets of various people on Twitter who use the tweet contains hate speech, the tweet is not offensive and the tweet uses offensive language but not hate speech.

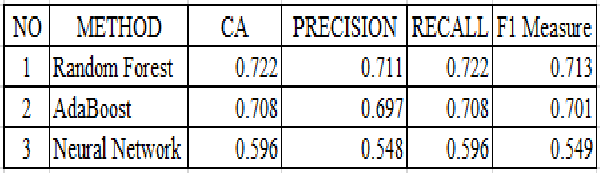
B. Preprocessing

The dataset of hate speech identification still needs to be processed in a preprocessing manner using the following methods:

1. Delete several characters such as: | :,; &! ? \
2. Normalize several hastags into the default word example is '#refugeesnotwelcome' become 'not welcome refugees'.
3. Giving lowercase letters and stemming processes,
4. Delete any tokens with document frequency less than 5
5. Delete the URL and username that are not needed.

C. Testing and Evaluation

Based on the classification results using Random Forest, Adaboost and Neural Network with stratified 10-fold cross validation the test results are obtained as follows:



NO	METHOD	CA	PRECISION	RECALL	F1 Measure
1	Random Forest	0.722	0.711	0.722	0.713
2	AdaBoost	0.708	0.697	0.708	0.701
3	Neural Network	0.596	0.548	0.596	0.549

Figure 5 : Experiment Result

The evaluation results of the 3 models used are Random Forest, AdaBoost and Neural Network, showing that the Random Forest model has a CA (Classification Accuracy) of 0.722, a precision of 0.711, a Recall of 0.722 and an F1 Measure of 0.713 which shows better test results compared with AdaBoost with an evaluation of CA (Classification Accuracy) of 0.722, Precision of 0.697, Recall of 0.708 and F1 Measure of 0.701 and Neural Network with CA (Classification Accuracy) of 0.596, Precision of 0.548, Recall of 0.596 and F1 Measure of 0.549 . From the comparison test data above it can be concluded that the Random Forest method performs a better prediction process compared to the AdaBoost method and the Neural Network uses the hatespeech identification dataset.

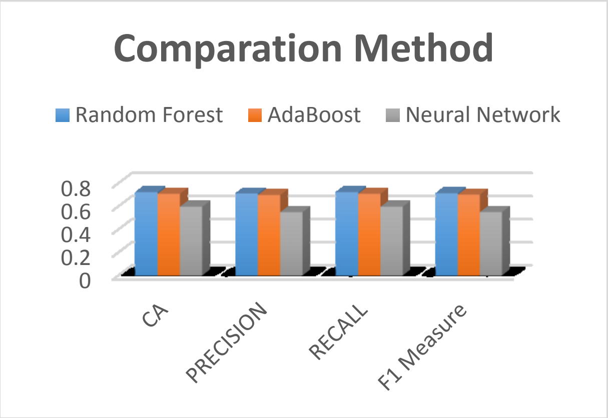
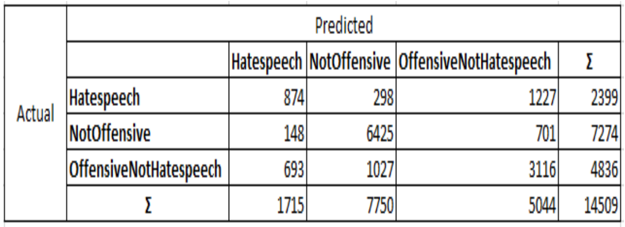


Figure 6 : Comparison chart using 3 methods

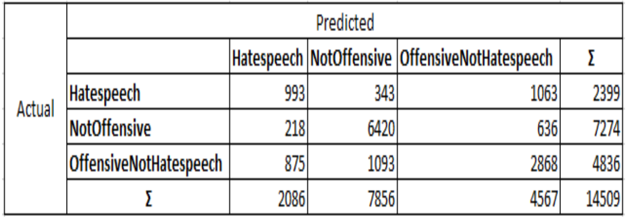
The results of the comparison of measurements of accuracy, precision, recall and F1 from cases of hate speech and crude speech are shown in the Graph above where Random Forest has better accuracy, precision and recall compared to Adaptive Booster and Neural Network. The evaluation results are also indicated by the table confussion matrix as follows:



		Predicted			
		Hatespeech	NotOffensive	OffensiveNotHatespeech	Σ
Actual	Hatespeech	874	298	1227	2399
	NotOffensive	148	6425	701	7274
	OffensiveNotHatespeech	693	1027	3116	4836
	Σ	1715	7750	5044	14509

Figure 7 : Confusion Matrix using Random Forest

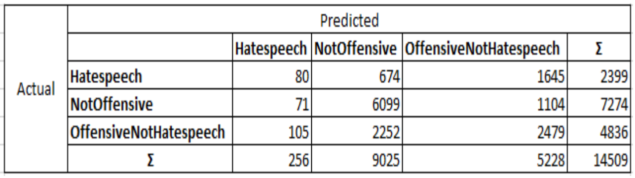
The figure 7 above shows that with the Random Forest method, there is a true prediction of 874 hatespeech, 6425 is not a offensive word and 3116 is a offensive word but not hatespeech with data of 14509 so that the prediction accuracy is 0.717.



		Predicted			
		Hatespeech	NotOffensive	OffensiveNotHatespeech	Σ
Actual	Hatespeech	993	343	1063	2399
	NotOffensive	218	6420	636	7274
	OffensiveNotHatespeech	875	1093	2868	4836
	Σ	2086	7856	4567	14509

Figure 8 : Confusion Matrix using AdaBoost

The figure 8 above shows that with the Adaboost method the correct prediction of 993 hatespeech is obtained, 6420 is not not a offensive word and 2868 is a offensive word but not hatespeech with data of 14509 so that the prediction accuracy is 0.708.



		Predicted			
		Hatespeech	NotOffensive	OffensiveNotHatespeech	Σ
Actual	Hatespeech	80	674	1645	2399
	NotOffensive	71	6099	1104	7274
	OffensiveNotHatespeech	105	2252	2479	4836
	Σ	256	9025	5228	14509

Figure 9 : Confusion Matrix using Neural Network

The figure 9 above shows that with the Neural Network method obtained correct predictions of 80 Hatespeech, 6099 not a offensive word and 2479 are offensive word but not Hatespeech with data of 14509 so that the prediction accuracy rate is 0.596.

V. CONCLUSION

Based on the test results on datasets about hate speech and offensive word using Random Forest, AdaBoost and Neural Network, it can be concluded that Testing 14509 datasets about hate speech and offensive word using Random Forest has a better level of accuracy and precision compared to the AdaBoost method and Neural Network, the results of testing using random forest also show that this method has a better level of recall compared to AdaBoost and Neural Network, Calculation of F1-Measure (F1) also shows that Random Forest has a higher value compared to AdaBoost and Neural Nework. After successfully detecting Hatespeech and the offensive word in the next study, the researchers will develop a method for detecting hatespeech and accents based on speech to get high accuracy in speech recognition.

REFERENCES

[1] W. Warner and J. Hirschberg, 'Detecting hate speech on the world wide web," *Proceeding LSM '12 Proc. Second Work. Lang. Soc. Media,* no. Lsm, pp. 19-26,2012.

[2] Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. Countering online Hate Speech. UNESCO.

[3] Ricardo Martins, Marco Gomes, Jos´e Jo˜ao Almeida, Paulo Novais, Pedro Henriques, Hate speech classification in social media using emotional analysis, IEEE 7th Brazilian Conference on Intelligent Systems,2018.

[4] Zewdie Mossie,Jenq-Haur Wang, Social Network Hate Speech Detection For Amharic Language, Dhinaharan Nagamalai et al. (Eds) : NATL, CSEA, DMDBS, Fuzzy, ITCON, NSEC, COMIT - 2018 pp. 41–55, 2018

[5] Ika Alfina,Rio Mulia, Mohamad Ivan Fanany,Yudo Ekanata, Hate Speech Detection in the Indonesian Language: A Dataset and Preliminary Study,International Conference On Advance Computer Science and Information System,2017.

[6] Edel Greevy and Alan F. Smeaton. 2004. Classifying Racist Texts Using a Support Vector Machine. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

[7] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. In Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust.

[8] Wang, W.; Chen, L.; Thirunarayan, K.; and Sheth, A. P. 2014.Cursing in english on twitter. In CSCW, 415–425.

[9] Hema Krishnan,M. Sudheep Elayidom,T. Santhanakrishnan,Emotion Detection of Tweets using Naïve Bayes Classifier,International Journal of Engineering Technology Science and Research,Volume 4, Issue 11,2017.

[10] Naufal Riza Fatahillah,Pulut Suryati,Cosmas Haryawan,Implementation Of Naive Bayes Classifier Algorithm On Social Media (Twitter) To The Teaching Of Indonesian Hate Speech,International Conference on Sustainable Information Engineering and Technology (SIET), 2017.

[11] Kelvin Kiema Kiilu, George Okeyo, Richard Rimiru, Kennedy Ogada,Using Naive Bayes Algorithm in detection of Hate Tweets,nternational Journal of Scientific and Research Publications, Volume 8, Issue 3, March 2018.

[12] Dr.R.L.K.Venkateswarlu, Dr. R. Vasantha Kumari, G.Vani JayaSri,Speech Recognition By Using Recurrent Neural Networks,International Journal of Scientific & Engineering Research Volume 2, Issue 6, June-2011 .

[13] M. Ali Fauzi,Random Forest Approach fo Sentiment Analysis in Indonesian Language,Indonesian Journal of Electrical Engineering and Computer Science,Vol. 12, No. 1, October 2018, pp. 46~50.

[14] Sanjana Sharma,Saksham Agrawal,Manish Shrivastava,Degree based Classification of Harmful Speech using Twitter Data,Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, pages 106–112.

[15] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. "Automated Hate Speech Detection and the Problem of Offensive Language." Proceedings of the 11th International Conference on Wneb and Social Media (ICWSM).

[16] Shervin Malmasi,Marcos Zampieri,Detecting Hate Speech in Social Media,Proceedings of Recent Advances in Natural Language Processing , pages 467–472,Varna, Bulgaria, Sep 4–6 2017.

[17] Jasdeep Singh Bhalla,Anmol Aggarwal,Using Adaboost Algorithm Along with Artificial Neural Networks for Efficient Human Emotion Recognition From Speech,2013 International Conference on Control Automation,Robotics and Embedded System(CARE).

[18] Ho, Tin Kam (1995). Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995.

[19] Breiman L (2001). "Random Forests". Machine Learning. 45 (1): 5–32. doi:10.1023/A:1010933404324.

[20] Hema Krishnan,M. Sudheep Elayidom,T. Santhanakrishnan,Emotion Detection of Tweets using Naïve Bayes Classifier,International Journal of Engineering Technology Science and Research,Volume 4, Issue 11,November 2017

[21] Freund, Y. and R. E. Schapire. 1997. A decision-theoretic generalization of online learning and an application to boosting. Journal of Computer and System Sciences 55(1): 119–139.