Communications and Network, **2012, 4, 189-195**
http://dx.doi.org/10.4236/cn.2012.43023 Published Online August 2012 (http://www.SciRP.org/journal/cn)

Gender Identification on Twitter Using the Modified Balanced Winnow

William Deitrick, Zachary Miller, Benjamin Valyou, Brian Dickinson, Timothy Munson, Wei Hu
Department of Computer Science, Houghton College, New York, USA
Email: Wei.Hu@houghton.edu

Received April 15, 2012; revised May 18, 2012; accepted June 16, 2012

ABSTRACT

With the rapid growth of web-based social networking technologies in recent years, author identification and analysis have proven increasingly useful. Authorship analysis provides information about a document's author, often including the author's gender. Men and women are known to write in distinctly different ways, and these differences can be successfully used to make a gender prediction. Making use of these distinctions between male and female authors, this study demonstrates the use of a simple stream-based neural network to automatically discriminate gender on manually labeled tweets from the Twitter social network. This neural network, the Modified Balanced Winnow, was employed in two ways; the effectiveness of data stream mining was initially examined with an extensive list of n-gram features. Feature selection techniques were then evaluated by drastically reducing the feature list using WEKA's attribute selection algorithms. This study demonstrates the effectiveness of the stream mining approach, achieving an accuracy of 82.48%, a 20.81% increase above the baseline prediction. Using feature selection methods improved the results by an additional 16.03%, to an accuracy of 98.51%.

Keywords: Gender Identification; Twitter; Modified Balanced Winnow; Neural Networks; Stream Data Mining; Feature Selection

1. Introduction

One of the largest areas of Internet growth in recent years has been social networking, which allows users to interact with others unconstrained by time or physical location. Many such services have appeared, with the most prominent including Facebook, MySpace, and Twitter. Since its launch in 2006, millions of users have joined Twitter.

The proliferation of social media exemplified by Twitter has sparked a great deal of research, and two common areas of interest have been age and gender. These studies are useful for marketing, advertising, and legal investigation [1]. This paper explores the use of the Modified Balanced Winnow Neural Network for identifying a Twitter user's gender. Twitter poses a unique challenge in the area of gender identification because of its compact style; tweets are limited to a length of 140 characters. Because of this brevity, it is common practice to use shortened forms of words, acronyms, and a wide range of emoticons in order to communicate using a minimal amount of space.

1.1. Gender Identification

Many studies in gender identification have focused on the link between gender and language. Each gender exhibits characteristic linguistic styles which have been observed in various disciplines. In the case of social media, researchers primarily make use of textual indicators, though username, profile description, and avatars may also be used. The normal lexicon of tweets includes vernacular language, colloquialisms formed to simplify writing, various stylistic devices created by Internet users, and URLs. These unique lexical additions include a massive proliferation of acronyms (LOL, BRB, etc.), "leet" speak, and emoticons [2]. "Leet" speak is a unique style of writing which found its origins on the Internet and involves the replacement of letters with numbers or other combinations of characters which result in a similar appearance (e.g. "1337" for leet). How frequently these forms of informal Internet language are used varies between social groups and gender groups.

The length of tweets has several implications for gender identification. Because tweets are limited to 140 characters, there is less content available to predict an author's gender. On the other hand, the character limit for tweets means that users must fit whatever they want to say into a smaller space. This has the effect of concentrating the user's writing style, increasing the necessity of using the characteristic text styles prevalent in social

media. Studies have shown that women have a tendency to make more frequent use of emotionally charged language, adjectives, adverbs, and apologetic language when compared with men, and that men tend to use more aggressive, authoritative language [2].

Due to the text-based nature of social media, and the high rate at which tweets are posted, algorithms attempting to perform gender identification have particular run-time requirements. To be practical, the algorithm should handle each new tweet individually as it comes in, with the results available for future use. To this end, this study employsa neural network using data stream mining techniques to discriminate gender of Twitter users.

1.2. Data Stream Mining

Most computational studies to date have used batch methods for gender identification, which in many cases expend considerable resources and time in exchange for increased accuracy [1]. For datasets which are incredibly large and constantly growing, this sacrifice of efficiency is unacceptable [3]. In these cases a stream approach may be used to evaluate the data in real time utilizing the natural flow of data as it enters the system [4]. This of course requires that instances be evaluated at a pace equal to or greater than the rate at which they arrive [5]. Despite the loss of accuracy resulting from making only a single pass, the ability to rapidly process enormous amounts of data makes stream mining preferable to batch-mining considering that millions of tweets are posted every day.

2. Data and Their Representations

Using the Twitter Streaming API, 36,238 tweets were collected. The lack of gender information on Twitter profiles necessitated manual classification of Twitter users. Accounts were filtered to remove all users where the gender could not be determined. These profiles include twitter accounts made for businesses, organizations, and advertisers. Any profile where the user does not speak English in all tweets was also removed as this study is only identifying the gender of English-speakers. After removing these users, one tweet from each account was selected, leaving a data set comprised of 3031 tweets.

Each tweet was then parsed and represented by a vector string based on 9170 features of three different types. The first type contained 95 1-gram character features. These features are the counts of the alpha-numeric characters found on a typical keyboard. The next 9025 features are the 2-gram counts based on the 95 individual characters. The 2-gram features capture some information about the structure of the text as they represent the counts of the occurrence of pairs. The last 50 features are the top distinguishing features defined in [1]. These features are used to determine some of the English-specific phrases and character sequences.

Once the 9170 features were extracted from the tweets, the tweets were split into two files in the Attribute Relation File Format (ARFF). The first file contained the training dataset of 1484 tweets, of which 939 were female-authored and 545 were male-authored. The training set was used for tuning the Modified Balanced Winnow as described in Section 3.3, and for feature selection as described in Section 2.1. The second file contained the testing dataset, with 1,547 tweets total (954 female, 593 male), and was used for the evaluation of the Modified Balanced Winnow classifier with and without feature selection applied.

Feature Selection

To accomplish a particular machine learning task, it is often useful to identify the most informative and relevant features. Feature selection reduces noise from irrelevant or inaccurate features and may even produce an increase in accuracy [6]. Preliminary tests on the tuning data set using only the top ten features (determined from the Symmetrical Uncertainty Algorithm) produced an accuracy of 67% versus 53% using the full feature set. A secondary benefit is the increase in speed. In this experiment the total feature set contains over 9000 features. Utilizing feature selection reduces this set to 53 features, which represents a significant run-time efficiency gain.

Seven different algorithms from the data mining utility WEKA were used to ensure that the most accurate and unbiased set of features was obtained [6]. The algorithms applied were Chi-Square, Information Gain, Information Gain Ratio, One-R, Relief, Symmetrical Uncertainty, and Filtered Attribute Evaluation, all of which use WEKA's "ranker" filter.

Chi-Square feature selection evaluates attributes individually by measuring the chi-squared statistic with respect to the other classes. Information Gain is synonymous with Kullback-Leibler divergence, utilizing a decision tree to calculate the entropy of a set of values. Information Gain Ratio is a slightly modified version of Information Gain. One-R uses just one parameter, specifically the minimum bucket size for discretization. The Relief Algorithm samples random instances and checks neighboring instances of the same and different class. Symmetrical Uncertainty is the measurement of correlation between two attributes, to determine which attributes have little inter-correlation.

3. Methods

3.1. Mistake Driven Online Learner

The Modified Balanced Winnow Neural Network is able to effectively process natural language tasks, making it

an excellent fit for gender identification. The Modified Balanced Winnow is a member of a family of neural networks called Mistake Driven Online Learners. These networks generate a learning model which is updated only when classification mistakes occur [7]. Formally, the Mistake Driven Online Learner creates a model w_0, generally represented in a matrix, then classifies incoming instances. If the classification is incorrect, w_0 is updated using a defined update rule. The pseudo-code for this type of neural networks is defined below.

Pseudo-Code for the Basic Mistake-Driven Online Learner

1) Initialize model w_0. Define function: $f(w_i, x_t)$
2) For $t = 1, 2, 3, \cdots, T$;
   a) Retrieve new example x_t
   b) Predict $\hat{y}_t = f(w_i, x_t)$ and compare it with actual class y_t.
   c) If $\hat{y}_t \neq y_t$:
      i) Update model $w_i \to w_{i+1}$.
   d) Else:
      i) Prediction was correct.

3.2. Balanced Winnow Neural Network

The Balanced Winnow Neural Network is a Mistake-Driven Online Learner that uses two models to classify instances. Because the Balanced Winnow classifier has two models u_i and v_i, it needs two separate update parameters to change them. The first parameter is defined such that $\alpha > 0$, which is a promotion factor. When a model is updated with α, the influence of this model is increased. The second parameter β decreases the magnitude of a model's influence and is defined in the range $0 < \beta < 1$. The Balanced Winnow algorithm also has a biasing parameter θ_{th} that serves as a threshold for classification. This parameter allows the Balanced Winnow to be further adjusted to fit the data optimally. The three parameters cause the Balanced Winnow algorithm to update the models and accurately determine the class of each instance [8,9].

The Balanced Winnow is based on two functions. The first is a scoring function that determines the predicted class. The scoring function is defined as

$$f = sign(x_t, u_i - x_t, v_i - \theta_{th})$$

where $sign(x)$ is the signum function, and (x_t, w_i) is the inner product between the current instance *x* and the current weight vector w_i. If the actual class of the instance is different from this score, then the second function is called. The second function updates the models using the α and β parameters and is defined below.

Update Rules for the Balanced Winnow Algorithm

Function: *updateModels*()

Given: Models u_i, v_i where u_i is the positive model and v_i is the negative, true class y_t

1) If $(y_t < 0)$:
   a) $u_{i+1} = u_i * \beta$,
   b) $v_{i+1} = v_i * \alpha$,
2) Else:
   a) $u_{i+1} = u_i * \alpha$,
   b) $v_{i+1} = v_i * \beta$.

3.3. Modified Balanced Winnow Neural Network

The Balanced Winnow is further improved with two changes to create the Modified Balanced Winnow Algorithm [7]. The first is the addition of *M*, which is a margin to ensure that the updates only occur when the prediction is completely correct. This causes the update function to be called when the true class of the instance is multiplied by the score function is less than or equal to *M*. The second difference is an improved Model Update function. The Modified Balanced Winnow multiplies the current model by the promotion and demotion parameters α and β and also by the incoming instance. This change can be seen below, with pseudo-code for the entire Modified Balanced Winnow also displayed.

3.3.1. Modified Update Rule: Allows the Instance to Have an Effect on the Model

Function: *updateModels*2()

Given: Models u_i, v_i where u_i is the positive model and v_i is the negative, instance y_t having true class y_t

1) If $(y_t < 0)$:
   a) $u_{i+1} = u_i * \beta * (1 - x_t)$,
   b) $v_{i+1} = v_i * \alpha * (1 + x_t)$,
2) Else:
   a) $u_{i+1} = u_i * \alpha * (1 + x_t)$,
   b) $v_{i+1} = v_i * \beta * (1 - x_t)$.

3.3.2. Full Pseudo-Code for the Modified Balanced Winnow Algorithm

1) Initialize models u_0 and v_0.
2) For $t = 1, 2, \cdots, T$:
   a) Receive new example x_t and add the bias.
   b) Normalize x_t.
   c) Calculate score function:
      i) $score = x_t, u_i - x_t, v_i - \theta_{th}$.
   d) Retrieve true class y_t.
   e) If $(score * y_t) \leq M$: //prediction is wrong
      i) Update Models2():
   f) Else
      i) Continue:

4. Results and Discussion

4.1. Results

4.1.1. Gender Classification Results with All Features

The Modified Balanced Winnow Algorithm was used to gauge the effectiveness of features extracted from the Twitter dataset. First, the performance of the Modified Balanced Winnow was measured using all 9170 features. Initial testing found an appropriate value of 1.5 for the *M* parameter. Parameters α and β required more extensive tuning to achieve optimal results. To expedite this process, a threaded tuner program capable of running many instances of the Balanced Winnow Algorithm was created and used to quickly try large ranges of α and β values. Using this technique, we found that $\alpha = 1.08$ and $\beta = 0.963$ give the best results, yielding an accuracy of 82.48%. A 20.81% increase was achieved over the baseline accuracy of 61.67% derived from predicting all users female. The improved performance demonstrates the effectiveness of the Modified Balanced Winnow Algorithm and the feature set used. The test results are summarized in **Table 1**, with the best results displayed in bold text.

4.1.2. Feature Selection Results

Using the ranker filter in the WEKA toolkit as described in Section 2.1, we selected 53 of the most relevant features from the 9170 features in the training dataset. Of these 53 features, four were 1-gram features, 47 were 2-gram features, and two were higher n-gram features proposed by [1]. The characters in the set of selected features included capital and lowercase letters, punctuation marks, numbers, and spaces. **Table 2** shows several of the selected features, as well as the number of times each of those features appeared in tweets posted by male and female authors and the total number of times each feature appeared. Note that underscores in the table's feature column represent space characters.

4.1.3. Gender Classification Results with Selected Feature Set

While the Modified Balanced Winnow Algorithm exhibited acceptable performance with the entire feature set, performance improved significantly when the 53 features were used. A detailed discussion of our feature selection results is given in Section 4.2. Using the 53 selected features the Modified Balanced Winnow Algorithm was able to achieve significantly higher accuracy, up to 98%. We found that the best α and β parameters for the Modified Balanced Winnow differed slightly from the parameters producing best performance for the full set of features, and thus followed the same tuning process described above for use with the full feature set. The results of this tuning process are shown in **Table 3**, with the best

Table 1. Modified balanced winnow performance metrics for various α and β values for full feature set.

Alpha	Beta	Acc	Prec	Sens	Spec	F-M
1.02	0.933	0.5953	0.4582	0.3052	0.7757	0.3664
	0.943	0.6303	0.5228	0.4064	0.7694	0.4573
	0.953	0.6736	0.5824	0.5245	0.7662	0.5519
	0.963	0.6522	0.5475	0.5346	0.7254	0.5410
1.04	0.933	0.6613	0.5638	0.5143	0.7526	0.5379
	0.943	0.6432	0.5355	0.5211	0.7191	0.5282
	0.953	0.6251	0.5109	0.5126	0.6950	0.5118
	0.963	**0.7944**	**0.7303**	**0.7352**	**0.8312**	**0.7328**
1.06	0.933	0.6542	0.5488	0.5497	0.7191	0.5493
	0.943	0.6542	0.5487	0.5514	0.7180	0.5500
	0.953	0.5966	0.4740	0.4772	0.6709	0.4756
	0.963	0.6303	0.5176	0.5211	0.6981	0.5193
1.08	0.933	0.6025	0.4815	0.4840	0.6761	0.4828
	0.943	0.6283	0.5151	0.5177	0.6971	0.5164
	0.953	0.6736	0.5738	0.5767	0.7338	0.5753
	0.963	**0.8248**	**0.7701**	**0.7740**	**0.8564**	**0.7721**

Table 2. Examples of the 53 selected features for gender identification (underscores represent space characters).

Feature	Male Count	Female Count	Total Count
_	5504	9360	14864
y_	240	498	738
my	43	120	163
ov	36	82	118
love	15	38	53
:)	3	48	51
oa	15	9	24
(:	0	13	13
?!	0	13	13
MA	9	2	11
1_	8	3	11
lb	9	1	10
mg	0	9	9
yn	7	1	8
DO	6	2	8
4_	7	0	7
Ll	4	0	4

Table 3. Modified balanced winnow performance metrics for various α and β values for selected feature set.

Alpha	Beta	Acc	Prec	Sens	Spec	F-M
1.1	0.8	0.958	0.9567	0.9325	0.9738	0.9445
	0.85	0.9632	0.9589	0.9444	0.9748	0.9516
	0.9	0.9748	0.9695	0.9646	0.9811	0.967
	0.95	**0.9774**	**0.9713**	**0.9696**	**0.9822**	**0.9705**
1.2	0.8	0.9664	0.9593	0.9528	0.9748	0.956
	0.85	0.9735	0.9662	0.9646	0.979	0.9654
	0.9	**0.9754**	**0.968**	**0.968**	**0.9801**	**0.968**
	0.95	0.9741	0.9663	0.9663	0.979	0.9663
1.3	0.8	0.9729	0.9646	0.9646	0.978	0.9646
	0.85	0.9716	0.9629	0.9629	0.9769	0.9629
	0.9	0.9729	0.9646	0.9646	0.978	0.9646
	0.95	**0.9851**	**0.9798**	**0.9815**	**0.9874**	**0.9806**
1.4	0.8	0.9716	0.9629	0.9629	0.9769	0.9629
	0.85	0.9722	0.963	0.9646	0.9769	0.9638
	0.9	0.9709	0.9613	0.9629	0.9759	0.9621
	0.95	**0.9877**	**0.9832**	**0.9848**	**0.9895**	**0.984**
1.5	0.8	0.9683	0.9579	0.9595	0.9738	0.9587
	0.85	**0.9748**	**0.9663**	**0.968**	**0.979**	**0.9671**
	0.9	0.9722	0.963	0.9646	0.9769	0.9638
	0.95	0.9716	0.9629	0.9629	0.9769	0.9629

results being displayed in bold text.

4.2. Discussion

4.2.1. Feature Selection Discussion

Many of the 53 features selected correlate to observations from other research regarding characteristics of gender-specific writing. For instance, the set of selected features included two features with digits, both of which occurred frequently in male-authored tweets [2]. This is consistent with other studies' findings stating that males tend to use specific quantities or values in their writing [2]. Our selected feature set also includes several features occurring predominantly in female-authored tweets. The emoticons ":)" and "(:" were both shown to be indicative of female authorship, and the features "love" and "ov" (primarily as part of the word "love") appeared mostly in tweets written by females as well. The feature "mg" also is strongly indicative of female authorship, appearing in the acronym "omg" in our dataset. These features are characteristic of strong emotion, corresponding to the fact that females produce more emotionally intense writing [2]. One additional characteristic of female authors is their tendency to use possessive pronouns [2], which is captured by the "my" feature in the selected feature set. To further illustrate the information captured by feature selection, **Table 4** lists several common words containing a subset of the selected features. Note that only features from **Table 2** without numeric or punctuation characters are shown, and several of the top words listed are proper nouns or Internet slang.

While using our selected features as input for the Modified Balanced Winnow Algorithm yielded significantly higher accuracy, several of the features selected are rather unintuitive. For instance, in the table above, "lb", and "Ll" seem particularly insignificant, and the words in which these features appear are uncommon proper nouns. However, the nature of our training and testing datasets gives insight as to why these features were selected. Both datasets were quite small considering the immense scope of Twitter, and were collected from the Twitter Streaming API over a relatively short period of time. As a result, it is not unreasonable that features specific to these relatively small datasets would exist within the set of selected features. In this case, the words "Bilbao", "Welbeck", and "Llorente" are all references to professional soccer, which was apparently a popular topic among male Twitter users when the training and testing datasets were collected. Thus, while these particular features would most likely not be helpful on other Twitter datasets, they are nevertheless aptly suited to this particular dataset and show the effectiveness of our feature selection technique.

4.2.2. Error Rates in Gender Identification with the Modified Balanced Winnow

To better understand the operation of the Modified

Table 4. Top words containing selected textual features.

Feature	Top Words
y_	my, they, really, by
my	my, myself, bummy, dummy, myth
ov	love, over, lovely, loves
love	love, lovely, loves, loved, lovee
oa	goal, keyboard, loads, skateboarding
MA	MADT, LMAO, LMAOO
lb	Bilbao, JailbreakCon, Welbeck
mg	omg
yn	tryna, sync, syncing
DO	DO, DONE, DOESN'T, DOWN, PARDON
Ll	Llorente

Balanced Winnow Algorithm, we generated histograms displaying the error rate as the algorithm is running with optimal parameters. We split the 1547 tweets from the training set into groupings of five and found the error rate for each. As can be seen in the histograms in **Figure 1**, the Modified Balanced Winnow Algorithmcommitted

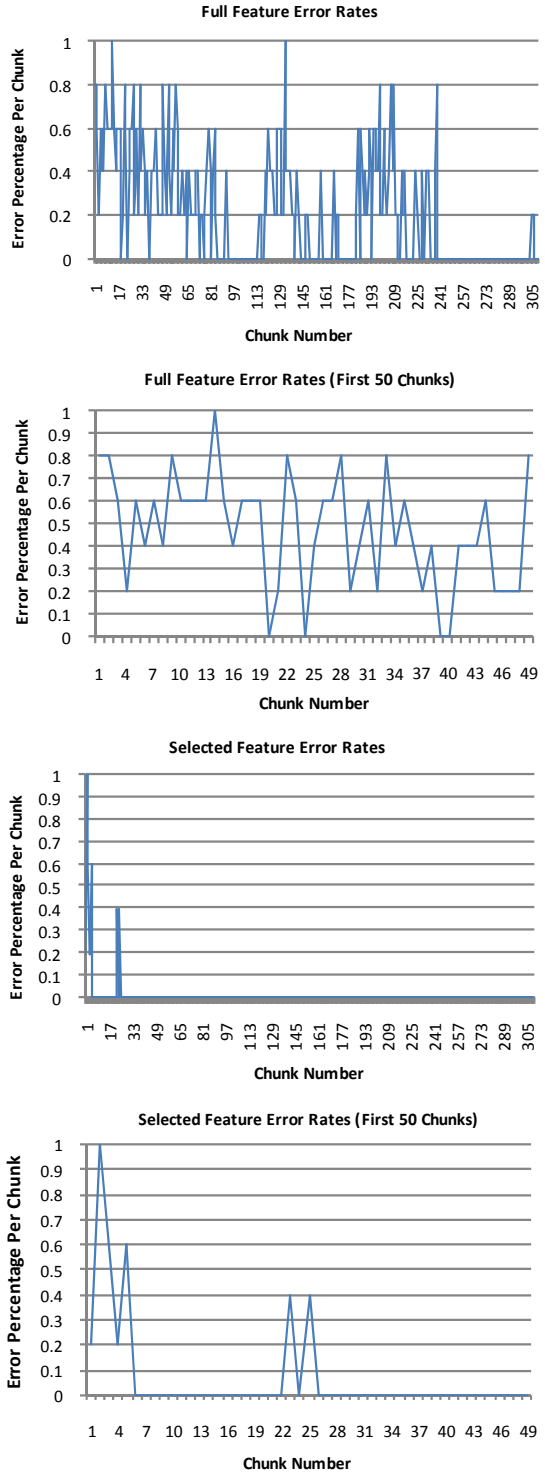

Figure 1. Balanced winnow error rates.

multiple mistakes in the first few chunks. When the selected features were tested, the algorithm had fewer errors. This means that, when using the small feature set, the Modified Balanced Winnow Algorithm has a very accurate initial model. In the first 50 groups, the selected feature tests only had two chunks with an error rate above 40%, while the full feature tests had roughly 40 sets above a 40% error rate. Also, after the initial 50 batches, the selected feature tests did not have any mistakes for the remainder of execution. This is further evidenced by the very high accuracy for the selected feature tests. In contrast with the high performance of evaluations using feature selection, tests performed on tweets using the full feature set were significantly less accurate. When using all 9170 features, the Balanced Winnow Algorithm continued to update its model, which can be attributed to an inaccurate initial model or the presence of features acting as noise. The error rates of the sets of five tweets allow us to visualize the effect particular tweets have on the neural network.

5. Conclusions

Gender detection of Twitter users is a complicated task. Because tweets are short and contain many slang words and frequent typos, it is difficult to find features informative enough to facilitate effective gender discrimination. In this study, we extracted 1-gram and 2-gram features and combined them with the published features in [1] to create 9170 features representing individual tweets. Using this entire set of 9170 features and a subset of 53 selected features, multiple tests were run using the Modified Balanced Winnow Algorithm.

The Modified Balanced Winnow Algorithm is a mistake-driven neural network that is well-suited to text analysis in a stream-mining environment. By tuning the α and β parameters of the Modified Balanced Winnow, we were able to achieve 82% accuracy with 77% precision using the entire set of features, and 98.51% accuracy with 97.98% precision using the 53 selected features. From these results we see that the Modified Balanced Winnow algorithm can effectively determine the gender of tweet authors, and that feature selection provides significant improvements in both accuracy and speed.

6. Acknowledgements

We would like to thank Houghton College for its financial support and for providing this research opportunity.

REFERENCES

[1] J. D. Burger, J. Henderson, G. Kim and G. Zarella, "Discriminating Gender on Twitter," Technical Report, Mitre

Corporation, Bedford.

[2] M. W. Corney, "Analyzing E-Mail Text Authorship for Forensic Purposes," Masters Thesis, Queensland University of Technology, Queensland, 2003.

[3] P. Refaeilzadeh, L. Tang and H. Liu, "Cross Validation," Arizona State University, 2008.

[4] W. Fan, H. Wang and P. S. Yu, "Active Mining of Data Streams," *Proceedings of the Fourth SIAM International Conference on Data Mining*, Florida, 22-24 April 2004.

[5] P. Domingos and G. Hulten, "Mining High Speed Data Streams," University of Washington, Washington DC, 2000.

[6] I. H. Witten and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques," 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[7] V. R. Carvalho and W. W. Cohen, "Single-Pass Online Learning: Performance," Voting Schemes and Online Feature Selection, KDD, 2006.

[8] I. Dagan, Y. Karov and D. Roth, "Mistake-Driven Learning in Text Cateorization," *Conference on Empirical Methods on Natural Language Processing*, 1997.

[9] O. N. Littlestone, "Learning Quickly When Irrelevant Attributes Abound: A New Linear-Threshold Algorithm," *Machine Learning*, Vol. 2, No. 4, 1988, pp. 285-318. doi:10.1007/BF00116827

