Gender Classification from Text

Chandana Roopa Reddy Rajala, Sakhitha Chowdary kanyadhara

Abstract— Classifying the blogs based on the gender of the author is the import application used by many other applications. Existing systems use the features like words, word classes and POS n-gram, POS sequence patterns. In this project, F measure, POS sequence pattern, bag of words and word classes. Words Classes are taken based on the psychology paper along and given by in [1]. Existing state-of-art methods are used to make a comparison study to find which methods retrieve the best features for better accuracy.

I. INTRODUCTION

Web blogs are like personal diaries and the language used in them is informal. Many companies use the blog information in rating theirs products or suggesting their products based on the author interest to market their products. The main difference between classifications of text of Reuters with that of blog data is the informal language used in blogs. The blogs are small and may many grammatical mistakes, abbreviations, phrases, wrong spelling and slangs

In this project, POS sequential patterns, F measure and class of words are used as features as proposed in [1]. Bag of words technique is used which collects the set of words depending on the occurrence of words in all the documents and these set of words are used as the features. It gives the information about the usage of parts-of-speech by the men and female. The bag-of-words model is a simplifying representation used in natural language processing and information retrieval (IR).To get our bags of words, we count the number of times each word occurs in each sentence. The number of words to be taken can be selected as required depending on the classification problem. Word classes conversation, home, family, food, romance, positive, negative and emotion are taken as features. The new word classes commonality, activity and collectivity are taken from [2].

Word2vec is a group of related models that are used to produce word embeddings. These models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words.

A POS sequence pattern is a sequence of consecutive POS tags that satisfy constraints. Using pos sequence patterns the authors writing style can be understood. Pos sequence patterns are taken as features for classification. POS n-gram helps in knowing the stylistic and syntactic features.

All the collected features are used for classification. SVM, SVM-R, Naive Bayes classifiers are used. The Principle component analysis, Chi-square and information gain are used for feature selection and to reduce the dimensionality of features.

II. RELATED WORK

Many papers have been published regarding the gender classification of text. Most of the papers These papers treat the problem as a classical machine learning one, and train a linear or Non-linear classifier using handcrafted, word-based features on a variety of textual sources. [11] Studied gender and text relationships in formal writing using the British National Corpus [11], and discovered that there was a clear difference in writing styles of male and female authors. [1] looked at the effectiveness of tweets to identify author gender. Their approach used n-grams concatenated with author's profile information to predict an author's gender, obtaining around 77% accuracy using texts alone, and notably above 90% accuracy with all features included. The current state-of-the-art using pure textual features was demonstrated by Mukherjee and Liu [1], who looked at content words, dictionary-based content analysis results, and POS (part-of-speech) tags for blog posts.

Their hand-crafted features consisted of the following categories:

1. Frequency Measure: Frequency of various parts of speeches. The F Measure is based upon the observation that males and females tend to have different preferences in frequency of types of parts of speech.
2. Stylistic Features: These features are characterized by the words used particularly in the blog context.
3. Gender Preferential Features: These features represent the tendency that females us more emotionally intensive adverbs and adjectives where as males tend to be more punctuated.
4. POS pattern features: These features represent patterns of POS tags.

III. DATASET

Blog Dataset that is taken has few flaws. Total given number of rows is 3232. But there are few empty text rows. The other thing we observed, it is said to be two labeled data, but there are 6 in total. ‘M’,’F’,’ M’,’ F’, m, f are the class labels. There are 127 ‘ M’s, 153 ‘ F’’s, 5 m’s, 4f’s. We processed the data and got 3215 rows to start the experiments.

IV. IMPLEMENTATION AND RESULT

A. *F-Measure*

We have taken the F-measure as a feature from [1]. The F-measure feature was originally proposed in (Heylighen and Dewaele, 2002). F-measure is different from F_Score. It gives the information about the usage of parts-of-speech by the men and female. It is shown that the F-score of male is greater than female writings, which indicate the preferring of men more formal writing. F-measure is defined based on the frequency of the POS usage in a text (*freq.x* below means the frequency of the part-of-speech *x*):

$$F = 0.5 * [(freq.noun + freq.adj + freq.prep + freq.art) - (freq.pron + freq.verb + freq.adv + freq.int) + 100]$$

First we have stored the dataset in a database. This is optional we can use even files. We used Stanford parser to tag the dataset. There are two models in the Stanford POS tagger. English-bidirectional-distim and English-left3words-distim. We used English-left3wods-distim model for this experiment. The tagset consists of 44 tags. These are called Penn Tree bank Tagset. After tagging all the reviews we processed the tagged data to find the words that end with NN, JJ, VB, . The next is counting the occurrence of the nouns, adjectives, prepositions, verbs, articles, adverbs and interjections in each document and updating the table. Finally this table is imported as a CSV file and we conducted experiments on this feature using many classification methods. We used Weka to perform different classifiers. Here as there are only one feature we did not implemented any kind of attribute selection methods. Instead we saved all the effort for the end. We used 10-fold and 5-fold cross validation for the naïve Bayes and Bagging. We did not considered this variation in folds because there is no significant change by doing so. But it gave the best results than all other classifiers. Bagging algorithm implemented here uses J48 tree instead of default option RepTree.

Classifier	Accuracy	F-Measure	Precision	Recall
Naive Bayes(10-fold)	56.8097	0.412	0.323	0.568
Naive Bayes(5-fold)	56.4677	0.426	0.505	0.565
SVM	56.8097	0.412	0.323	0.568
Bagging(J48-5fold)	57.5249	0.496	0.560	0.575
Bagging(J48-10fold)	57.8358	0.491	0.570	0.578

Table1 Results for F-measure as a feature

B. *Bag-of-words*

The bag-of-words model is a simplifying representation used in natural language processing and information retrieval (IR). In this model, each review in the dataset is represented as the bag of its words, disregarding grammar and even **word** order but keeping multiplicity. It does not have any ordered representation. The only thing that matter is the count of the words. The Bag of Words model learns a vocabulary from all of the documents, and then models each document by counting the number of times each word appears.

To get our bags of words, we count the number of times each word occurs in each sentence. To limit the size of the feature vectors, we should choose maximum vocabulary size as 1000. This 1000 is the most frequent words used in the dataset. Before doing this one we need to remove all the stop words like is, are, this, that. Because they may not give significant results but there are frequently used. So such kind of words is to be removed for more efficient output. We have used Vectorizer to create bag of words. Here there is a parameter called maxfeatures, this we have taken it for 400, 500 and 1000 most frequent words. After applying different classifiers like naives bayes, SMO(sequential algorithm) algorithm that with poly kernel and default c=1 and default learning parameter as 0.01. Taking into500 features gave maximum accuracy of approximately 65%. But taking this amount of features for final classification did not give likely results. So we settled with 1000 features.

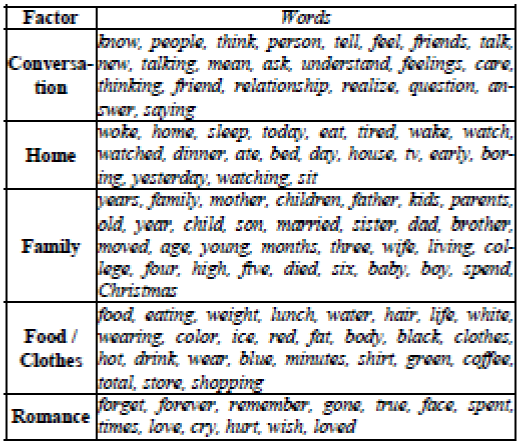
First we wrote a program to extract the 1000 features and saved them to an CSV file for future use. The resultant file has the values in scientific format. We converted them to a numeric format. We used Weka to perform different classifiers. Here as there are only 1000 features we did not implemented any kind of attribute selection methods. Instead we saved all the effort for the end. We used 10-fold and 5-fold cross validation for the naïve Bayes and SVM. RandomForest is a tree that took a lot of time compared to all the classifiers used to classify the data. But it gave the best results than all other classifiers. Bagging algorithm implemented here uses J48 tree instead of default option RepTree.

Classification	Accuracy	Precision	Recall	F-Measure
Bagging	61.5552	0.626	0.616	0.615
SMO(10-fold)	62.6439	0.634	0.626	0.615
SMO(5-fold)	61.9596	0.626	0.620	0.609
Naïve Bayes(5-fold)	60.8709	0.609	0.609	0.609
Naïve Bayes(10-fold)	61.8974	0.619	0.619	0.619
RandomForest	63.7325	0.637	0.637	0.635

Table2 Results for Bag-of-words Model

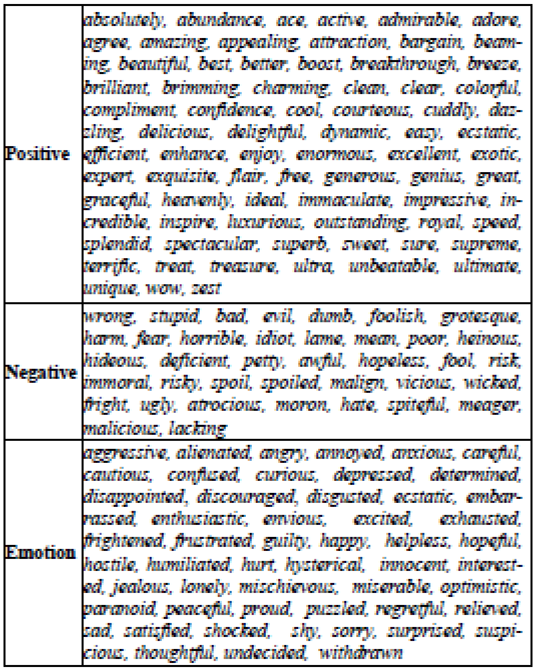
C. *Word classes*

Word classes’ conversation, home, family, food, romance, positive, negative and emotion are taken as features.



Factor	*Words*
Conversation	*know, people, think, person, tell, feel, friends, talk, new, talking, mean, ask, understand, feelings, care, thinking, friend, relationship, realize, question, answer, saying*
Home	*woke, home, sleep, today, eat, tired, wake, watch, watched, dinner, ate, bed, day, house, tv, early, boring, yesterday, watching, sit*
Family	*years, family, mother, children, father, kids, parents, old, year, child, son, married, sister, dad, brother, moved, age, young, months, three, wife, living, college, four, high, five, died, six, baby, boy, spend, Christmas*
Food / Clothes	*food, eating, weight, lunch, water, hair, life, white, wearing, color, ice, red, fat, body, black, clothes, hot, drink, wear, blue, minutes, shirt, green, coffee, total, store, shopping*
Romance	*forget, forever, remember, gone, true, face, spent, times, love, cry, hurt, wish, loved*

Table 3 Word Factors I



Positive	*absolutely, abundance, ace, active, admirable, adore, agree, amazing, appealing, attraction, bargain, beaming, beautiful, best, better, boost, breakthrough, breeze, brilliant, brimming, charming, clean, clear, colorful, compliment, confidence, cool, courteous, cuddly, dazzling, delicious, delightful, dynamic, easy, ecstatic, efficient, enhance, enjoy, enormous, excellent, exotic, expert, exquisite, flair, free, generous, genius, great, graceful, heavenly, ideal, immaculate, impressive, incredible, inspire, luxurious, outstanding, royal, speed, splendid, spectacular, superb, sweet, sure, supreme, terrific, treat, treasure, ultra, unbeatable, ultimate, unique, wow, zest*
Negative	*wrong, stupid, bad, evil, dumb, foolish, grotesque, harm, fear, horrible, idiot, lame, mean, poor, heinous, hideous, deficient, petty, awful, hopeless, fool, risk, immoral, risky, spoil, spoiled, malign, vicious, wicked, fright, ugly, atrocious, moron, hate, spiteful, meager, malicious, lacking*
Emotion	*aggressive, alienated, angry, annoyed, anxious, careful, cautious, confused, curious, depressed, determined, disappointed, discouraged, disgusted, ecstatic, embarrassed, enthusiastic, envious, excited, exhausted, frightened, frustrated, guilty, happy, helpless, hopeful, hostile, humiliated, hurt, hysterical, innocent, interested, jealous, lonely, mischievous, miserable, optimistic, paranoid, peaceful, proud, puzzled, regretful, relieved, sad, satisfied, shocked, shy, sorry, surprised, suspicious, thoughtful, undecided, withdrawn*

Table 4 Word Factors II

New class of words are taken into account according to study in [2]. This study considers language used by the female and male blog users and developed three word classes certainty, activity and commonality.

Certainty is calculated as: [Tenacity + Leveling Terms + Collectivity + Insistence] - [Numerical Terms + Ambivalence + Self-Reference + Variety]

- Where *Tenacity:* Words that express confidence and totality.

Leveling: Words that express completeness and assurance. *Collectivity:* Words that express social groupings such as a crowd or a world. The value of collectivity is the occurrence of all the words related to collectivity

.

- *Numerical Terms:* Any date or number that deals with quantitative or numerical operations. The value of numerical terms is the occurrence of all numerical
- *Ambivalence:* Words that express hesitation or uncertainty. This includes hedge phrases, vagueness, or confusion. The value of ambivalence is the occurrence of all the words related to ambivalence
- *Self-Reference:* Words that express first-person references, such as I, I'd, I'll, me, my, mine. The value of self-reference is the occurrence of all the words related to self-reference.

Activity:
Activity is calculated as: [Aggression + Accomplishment + Communication + Motion] - [Cognitive Terms + Passivity + Embellishment]. Each of these language scores is defined as follows:

- *Aggression:* Words that express competition or forceful action, including terms that imply physical energy or domination. The value of aggression is the occurrence of all the words related to aggression
- *Accomplishment:* Words that express the completion of a task, or methodical human behavior. The value of accomplishment is the occurrence of all the words related to accomplishment
- *Communication:* Words that express social interaction, including face-to-face or mediated modes, such as a film or telephone. The value of communication is the occurrence of all the words related to communication
- *Motion:* Words that express movement, speed, journeys, transit, or physical processes. The value of motion is the occurrence of all the words related to motion
- *Cognitive Terms:* Words that express “cerebral processes”, including discovery, psychology, logic, mental challenges, or learning practices. The value of cognitive is the occurrence of all the words related to cognitive
- *Passivity:* Words that express inactivity, compliance, or docility. The value of passivity is the occurrence of all the words related to passivity

Commonality: Subcomponents and Calculation
Commonality is calculated as: [Centrality + Cooperation + Rapport] - [Diversity + Exclusion + Liberation]. Each of these language scores is defined as follows:

- *Centrality:* Words that express regularity, congruence, predictability, universality, or an agreement on central values. The value of centrality is the occurrence of all the words related to centrality
- *Cooperation:* Words that express formal, informal, and intimate associations and interactions. The value

of cooperation is the occurrence of all the words related to cooperation.

- *Rapport:* Words that express an affinity toward similarities among a group of people. The value of rapport is the occurrence of all the words related to rapport.
- *Diversity:* Words that express non-conformity or heterogeneity. The value of tenacity is the occurrence of all the words related to diversity.
- *Exclusion:* Words that express social isolation. The value of exclusion is the occurrence of all the words related to exclusion.
- *Liberation:* Words that express a rejection of social standards. The value of liberation is the occurrence of all the words related to liberation.

Centrality	Axis, pivot, interior, midpoint, place, focus, regular, congruence
Cooperation	collaboration, joint action, combined effort, teamwork, partnership, coordination, liaison, association, synergism, give and take, compromise
Rapport	affinity, close relationship, understanding, mutual understanding, bond, empathy, sympathy, accord
Diversity	variety, miscellany, assortment, mixture, mix, mélange, range, array
Exclusion	Exclude, isolate
Liberation	Emancipation, enfranchisement, deliverance, disengagement, disenthrall

Table 5: Word factors for Commonality

The frequency of the words related to the words in the word class must be calculated to find the overall value of the feature. Wordnet is used to find the synset of words. Using this all the related words are found and kept in a list and their count in each document is calculated. And their sum is taken as the term frequency for that word in the word class. In the same way term frequency for all the related words in the word class is calculated and their sum is value of the word in the word class.

Similar way values for all the words in the word class are calculated and substituted in the formula given to get the overall feature value.

Each word class is taken as feature and the value of the feature is taken as occurrence of all the words in class in each document.

All the words in a word class are kept in a list and frequency of each word in the word class is calculated and the sum of all the words is taken as the value of that feature i.e. word class. In the similar way the values for all the word classes are calculated.

Classification	Accuracy	Precision	Recall	F-Measure
Bagging	50.1089	0.497	0.501	0.494
SMO(10-fold)	52.0373	0.271	0.520	0.356
Naïve Bayes (10-fold)	48.2426	0.494	0.482	0.417
RandomForest	49.0824	0.489	0.491	0.489

Table 6: Results for Word Class as features

D. *Word2Vec*

Word2vec is a group of related models that are used to produce word embeddings. These models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words.

- Taking context into consideration may lead to valid results.
- The dataset is modeled such that for the given dataset, the number of features yielded is 300, window_size=5,context=10,minimum count word is 30.
- Performed RandomForest classifier for the resultant features gave 64.3% accuracy.

E. *POS sequence patterns*

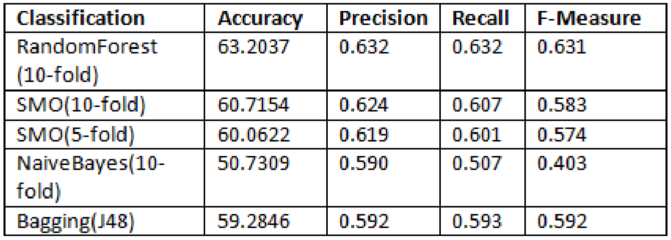
A POS sequence pattern is a sequence of consecutive POS tags that satisfy constraints. Using pos sequence patterns the authors writing style can be understood.

Pos sequence patterns are taken as features for classification. POS n-gram helps in knowing the stylistic and syntactic features. POS sequence pattern algorithm finds the sequence of POS tags of flexible length which satisfies the minimum support and minimum adherence.

POS Sequence algorithm:
The POS n-gram tagging is done on the training data.
Input: pos tagged training data, pos tag set, minimum support and adherence
Output: pos sequence patterns
1. candidate-gen() function is applied to generate patterns of length k. Depending on the minimum support and minimum adherence those generated patterns are discarded. The patterns that satisfy the minimum support and minimum adherence are added to the sequence pattern set.
2. The support of each sequence pattern is calculated by finding the probability of the pattern in the data set given i.e. the sequence pattern is checked for its presence in all the documents.
3. Adherence of a sequence pattern measures the strength of POS tags in sequence.
4. candidate-gen() function adds each POS n-gram tag before the existing sequence patterns.
5. candidate-gen() is iteratively applied until we obtain the sequence patterns of length four.

The training dataset is imported and the POS tagging is done to all the documents using NLTK POS tagging. List of POS tags and sequential patterns are also taken. The sequential pattern list initial has all the POS tags. The sequence patterns of length two are formed by appending sequential patterns with POS tag list and then minimum support and minimum adherence for all those sequence patterns are checked. If they are satisfied then it is added to the sequential pattern list. The same process is repeated for the patterns of length 3, 4 in the sequential order to obtain the list of sequential patterns which are taken as features. The feature value is the occurrence of that pattern in a document.

After finding all the possible best patterns in a file, we saved them to a file. These patterns are included to the CSV file that has all other features. But below are the classification results for just the patterns as the features using weka. There are 225 such patterns.



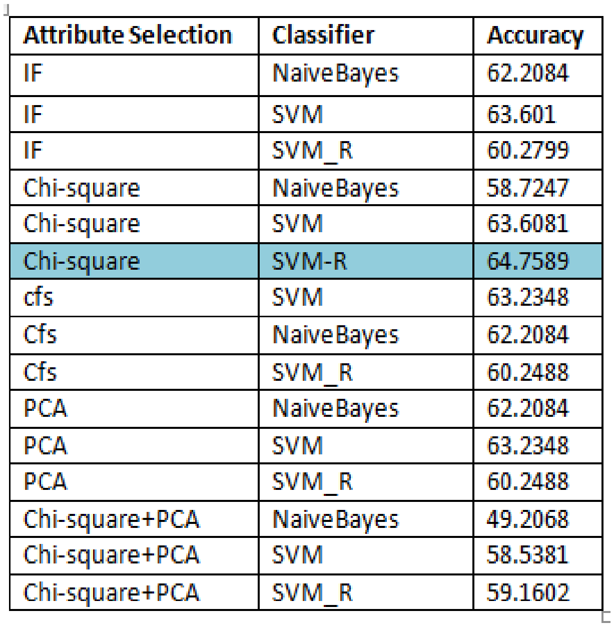
Classification	Accuracy	Precision	Recall	F-Measure
RandomForest (10-fold)	63.2037	0.632	0.632	0.631
SMO(10-fold)	60.7154	0.624	0.607	0.583
SMO(5-fold)	60.0622	0.619	0.601	0.574
NaiveBayes(10-fold)	50.7309	0.590	0.507	0.403
Bagging(J48)	59.2846	0.592	0.593	0.592

Table 7: Results for POS sequence pattern

V. Results and Discussion

We included all the features together in one CSV file. There are 1000 features for bag-of-words model, 153 word classes taken from [1] and one F_measure feature. Other three word class that was explained. The results are not upto the mark as expected. May be this is because the reference journal [2] taken is concentrated on teen bloggers. But we assumed it is applicable to all the age categories. The current dataset has the male and female of all the ages and not just the teenagers. The topics are related to many different sectors. Some are about their work, friends, completely formal, spiritual. This is the reason we think as a reason for the results. We also observed that the three word factors considered as very sparse data. To get the most important features among the features we got, we implemented three attribute selections information gain, chi-square, PCA and cfs. We applied all these attribute selection methods to all the classifiers SVM classification, SVM Regression and naïve bayes.

We found that by using chi-square attribute selection the accuracy is comparatively greater than other techniques. This is also applicable for only SVM using regression.



Attribute Selection	Classifier	Accuracy
IF	NaiveBayes	62.2084
IF	SVM	63.601
IF	SVM_R	60.2799
Chi-square	NaiveBayes	58.7247
Chi-square	SVM	63.6081
Chi-square	SVM-R	64.7589
cfs	SVM	63.2348
Cfs	NaiveBayes	62.2084
Cfs	SVM_R	60.2488
PCA	NaiveBayes	62.2084
PCA	SVM	63.2348
PCA	SVM_R	60.2488
Chi-square+PCA	NaiveBayes	49.2068
Chi-square+PCA	SVM	58.5381
Chi-square+PCA	SVM_R	59.1602

Table 8: Result for all features using different attribute selection and classifiers

VI. CONCLUSION

Existing state-of-art method use the features like words, word classes and POS n-gram, POS sequence patterns. In this project, F measure, POS sequence pattern, bag of words and word classes taken based on the psychology paper along with the word classes given by in [1]. All these are used as features and existing state-of-art methods are used to find the best features for better accuracy. But the problems with the new word classes were discussed above. Accuracies for information gain, Chi-squares and PCA were discussed above. PCA gave the least accuracies.

REFERENCES

[1] *Arjun Mukherjee and Bing liu for Improving Gender Classification of Blog Authors*

[2] *http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2005.tb00238.x/full.*

[3] Aric Bartle, Jim Zheng, for Gender Classification using deep learning.

[4] Agrawal, R. and Srikant, R. 1994. *Fast Algorithms for* Mining Association Rules. VLDB. pp. 487-499.

[5] Argamon, S., Koppel, M., J Fine, AR Shimoni. 2003.*Gender, genre, and writing style in formal written texts*. Text-Interdisciplinary Journal, 2003.

[6] Argamon, S., Koppel, M., Pennebaker, J. W., Schler, J. 2007. *Mining the Blogosphere: Age, Gender and the varieties of self-expression*, First Monday, 2007 - firstmonday.org

[7] Baayen, H., H van Halteren, F Tweedie. 1996. *Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution*, Literary and Linguistic Computing, 11, 1996.

[8] Blum, A. and Langley, P. 1997. *Selection of relevant features and examples in machine learning*. Artificial Intelligence, 97(1-2):245-271.

[9] BookBlog, *Gender Genie*, Copyright 2003-2007, http://www.bookblog.net/gender/genie.html

Borgelt, C. 2003. *Bayes Classifier Induction*. http://www.borgelt.net/doc/bayes/bayes.html

[10] Chung, C. K. and Pennebaker, J. W. 2007. *Revealing people's thinking in natural language: Using an automated meaning extraction method in open–ended self–descriptions*, J. of Research in Personality.

[11] Corney, M., Vel, O., Anderson, A., Mohay, G. 2002. *Gender Preferential Text Mining of E-mail Discourse*. 18th annual Computer Security Applications Conference (ACSAC), 2002.

[12] J. Dean and S. Ghemawat. 2004. *Mapreduce: Simplified data processing on large clusters*, Operating Systems Design and Implementation, 2004.

Forman, G., 2003. *An extensive empirical study of feature selection metrics for text classification.* JMLR,

[13] Rogati, M. and Yang, Y.2002. *High performing and scalable feature selection for text classification*. In CIKM, pp. 659-661, 2002.

[14] Schiffman, H. 2002. Bibliography of Gender and Language. http://ccat.sas.upenn.edu/~haroldfs/ popcult/ bibliogs/gender/genbib.htm

[15] Schler, J., Koppel, M., Argamon, S, and Pennebaker J. 2006. *Effects of age and gender on blogging*, In Proc. of the AAAI Spring Symposium Computational Approaches to Analyzing Weblogs.

[16] Silva, J., Dias, F., Guillore, S., Lopes, G. 1999. *Using LocalMaxs Algortihm for the Extraction of Contiguous and Noncontiguous Multiword Lexical Units*. Springer Lecture Notes in AI 1695, 1999

[17] Srikant, R. and Agrawal, R. 1996. *Mining sequential patterns: Generalizations and performance improvements* In Proc. 5th Int. Conf. Extending Database Technology (EDBT'96), Avignon, France.

[18] Tannen, D. (1990). You just don't understand, New York: Ballantine.

[19] Tsuruoka, Y. and Tsujii, J. 2005. *Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data*, HLT/EMNLP 2005, pp. 467-474.

[20] Tuv, E., Borisov, A., Runger, G., and Torkkola, K. 2009. *Feature selection with ensembles, artificial variables, and redundancy elimination.* JMLR, 10.

[21] Yan, X., Yan, L. 2006. Gender Classification of Weblog *Authors*. Computational Approaches to Analyzing Weblogs, AAAI.