Gender Clustering of Blog Posts using Distinguishable Features

Yaakov HaCohen-Kerner, Yarden Tzach and Ori Asis
Department of Computer Science, Jerusalem College of Technology (Machon Lev),
21 Havaad Haleumi St., P.O.B. 16031, 9116001 Jerusalem, Israel

Keywords: Blog Posts, Distinguishable Features, Gender Clustering.

Abstract: The aim of this research is to find out how to perform effective clustering of unlabeled personal blog posts written in English by gender. Given a gender-labeled blog corpus and a blog corpus that is not gender-labeled, we extracted from the labeled corpus distinguishable unigrams for both males and females. Then, we defined two general features that represent the relative frequencies of the distinguishable males' unigrams and females' unigrams, (males' frequency and females' frequency). The best distinguishable feature was found to be the males' frequency feature with a ratio factor at least 1.4 times that of females. This feature leads to accuracy rate of 83.7% for gender clustering of the unlabeled blog corpus. To the best of our knowledge, this study presents two novelties: (1) this is the first study to cluster blog posts by gender, and (2) clustering of an unlabeled corpus using distinguishable features that were extracted from a labeled corpus.

1 INTRODUCTION

Due to the explosion of information on the Internet and its increased availability there is a need to automatically perform text classification. However, most of the texts are not pre-assigned to classes, and therefore they are unsuitable for supervised machine learning (ML). Hence, document clustering using unsupervised ML is necessary.

Clustering is an automatic grouping of unlabeled text documents into groups, which are called clusters. Clustering of documents is the process of creating a set of clusters in such a way that documents within one cluster are more similar and documents from different clusters are more dissimilar (Tryon, 1939; Bailey, 1994). Clustering is applied in various domains including bioinformatics (Tasoulis et al., 2004), data mining (Fayyad et al., 1996), genetics (Shamir and Sharan, 2000), machine vision (Cucchiara, 1998), and social sciences (Aldenderfer and Blashfield, 1984).

The research presented in this paper was performed in the blog domain. "Blog is a popular and flexible way to publish information and express feelings, especially for private use" (Gao and Lai, 2010). The selected application domain is blog posts clustering by gender. The motivation for gender classification and clustering has grown during the last years, with rise of the digital age and the increase in human-computer interaction (Ngan and Grother, 2015).

Furthermore, the language used by an author is impacted by variables such as the author's age and gender (Eckert, 1997; Eckert and McConnell-Ginet, 2013). Bucholtz and Hall (2005) showed that speakers use language as a resource to construct their identity. In many cases, the person's gender identity can be identified by finding the linguistic features associated with male or female speech. These linguistic features gain social meaning in a cultural and societal context. On Twitter, for instance, users construct their identity through interacting with other users (Marwick and Boyd, 2011).

The main problem in the gender clustering task is identifying the clustering properties, which can be clearly distinguished from one cluster to another and decide how clusters should be defined. The task of clustering by gender is sometimes difficult even for a person to perform. The gender of the author of a blog can be conjectured based on the subjects discussed or its writing style (Schler et al., 2006; Schwartz et al., 2013). However, even then there is no promise for perfect identification in all cases. Sometimes females write about topics that are considered as masculine subjects (e.g., computers, electronics, and politics) or use male writing style

HaCohen-Kerner, Y., Tzach, Y. and Asis, O.
Gender Clustering of Blog Posts using Distinguishable Features.
DOI: 10.5220/0006077403840391
In *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016) - Volume 1: KDIR*, pages 384-391
ISBN: 978-989-758-203-5

and sometimes males write about topics that are considered as feminine subjects (e.g., home, family, and feelings) or use female writing style (Schler et al., 2006). One of the most basic methods to differentiate between males and females is to have two word lists so that one list includes words that are relatively more common for males and the second list includes words that are relatively more common for females.

We worked with two corppra that are of the same type, personal blogs written in English. The first corpus is from August 2004 and the second corpus is from 2012. The blogs of the first corpus are already gender-labeled while the blogs of the second corpus are not gender-labeled.

To the best of our knowledge, our research is the first to cluster blog posts by gender. Blog clustering by gender is important due to the huge number of unlabeled texts, which is available in the Web in general and in blogs in particular. Automatic gender clustering will enable people and companies to learn new things about females versus males and to take advantage of this knowledge (e.g., for marketing purposes).

This study claims that blog posts clustering using distinguishable features extracted from a labeled corpus lead to better clustering results for another similar corpus, which is not labeled. That is to say, we efficiently perform gender clustering on an unlabeled corpus using features extracted from a corpus, which is labeled. We did not see any use of such distinguishable features in former relevant clustering studies.

The main contribution of this study is the presentation of distinguishable unigrams that were extracted from a labeled blog corpus and were found as successful features for clustering of a totally different and unlabeled blog corpus from the same domain, personal blog posts written in English. We found that males distinguishable unigrams with a ratio factor at least 1.4 times that of females lead to an accuracy rate of 83.7% for gender clustering. The use of other features (from other feature sets and/or from the same feature set) harmed the clustering results.

This paper is organized as follows: Section 2 gives an overview for a much related domain, gender classification, Section 3 describes blog clustering and suitable feature sets. Section 4 introduces the selected features for this research. Section 5 presents the clustering model. Section 6 describes the experimental results and their analysis. Finally, Section 7 presents a summary and proposals for research directions.

2 GENDER CLASSIFICATION

A related domain to gender clustering is gender classification. Gender classification in Natural Language Processing (NLP) is the supervised learning task of assigning natural language text documents to males or females according to their content. Author gender identification has been studied both as an authorship attribution task and gender classification task (Koppel et al., 2002).

Koppel et al. (2002) showed that automated text classification techniques can exploit combinations of simple lexical and syntactic features to infer the gender of the author of an unseen formal written documents (writing styles in modern English books and articles). The best accuracy results (around 80%) have been obtained when using both function words and parts-of-speech n-grams.

Yan and Yan (2006) constructed a corpus containing 75,000 individual blog entries authored by 3000 bloggers. They also presented a Naïve Bayes classification approach to identify genders of weblog authors. In addition to features employed in traditional text classification, the authors used weblog-specific features, e.g., web page background colors and emoticons. They presented the most "gender-discriminant" unigrams that they have found, e.g., "hit", "man", "peace", "played", and "yo".

Mukherjee and Liu (2010) proposed two novel methods to improve the state-of-the-art accuracy results. Their first method introduces a new class of features, which are variable length POS sequence patterns mined from the training data using a sequence pattern mining algorithm. The second method is a new feature selection method, which is based on an ensemble of several feature selection criteria and approaches. Empirical evaluation using a real-life blog data set shows that these two methods significantly improve the classification accuracy of the current state-of-the-art methods.

Burger et al. (2011) constructed a large, multilingual dataset labeled with gender and presented a few configurations of a language-independent classifier for identifying the gender of Twitter users. The best classifier performed at 92% accuracy.

Filho et al. (2016) performed gender classification by using 60 textual meta-attributes for the extraction of gender expression linguistic cues in tweets written in Portuguese. The authors take into account a widespread variety of features: characters, syntax, words, structure and morphology. They classified free texts posted on Twitter according to

author's gender using three different ML algorithms as well as evaluate the influence of the proposed meta-attributes in this process.

3 BLOG CLUSTERING AND SUITABLE FEATURE SETS

As mentioned in Section 1, clustering is an automatic grouping of unlabeled text documents into groups, which are called clusters. In general, text clustering is much less popular than text classification (can be proved by numbers of papers and citations), which is an automatic grouping of labeled text documents into groups. There are significantly much more studies about text classification than about text clustering. The same (clustering versus classification) is true for gender and/or blog tasks. Probably, the main reasons for these findings are that: (1) text clustering is much harder to evaluate than text classification and (2) there are much less clustering (unsupervised ML) methods compared to supervised ML methods. Therefore, the accuracy results achieved by text clustering tasks are usually significantly lower than the results achieved by text classification tasks.

Clustering of blogs based on similar content using blog tags was performed by Brooks and Montanez (2006). They used the top 350 blog tags and they found that the tags are useful for clustering of articles into broad clusters, but less effective in indicating the particular content of an article. They showed that automatic extraction of words deemed to be highly relevant lead to better categorization of articles. Kuzar and Navrat (2011) present Slovak blog clustering enhanced by comments of web users. They combined content clustering with implicit ties between users based on comments. According to the results of their experiments, the quality of content clusters can be improved by considering implicit ties between commentators in case of articles, which do not fit into a single cluster.

Various feature sets have been applied in clustering tasks. Miao et al. (2005) applied three types of feature sets for document clustering: (1) words (after removing stopwords, stemming, pruning rare terms, and tf-idf weighting), (2) terms (based on their C Value, i.e., a frequency-based weight that accounts for nested terms), and (3) frequent character n-grams. They found that the n-gram-based representation provides the best results. Banerjeeet al. (2007) introduced a method of improving the accuracy of clustering short texts by enriching their representation with additional features from Wikipedia. Empirical results showed that their enriched representation of text items substantially improved the clustering accuracy when compared to the conventional bag of words representation. HaCohen-Kerner and Margaliot (2013; 2014) proposed for clustering various types of word unigrams, e.g., most frequent words (FW) including function words (stopwords), most frequent filtered words (FFW) excluding function words, and words with the highest variance values (HVW).

Nguyen et al. (2014) explained why it is hard to predict gender and age from tweets. They showed that most research so far treats gender and age as fixed variables and ignores that language use is related to the social identity of speakers, which may be different from their biological identity. In their research, they showed that approaching age and gender as social variables allow for richer analyses and more robust systems.

4 SELECTED FEATURES

In this research, we consider until now features belong to only two feature sets: 45 Parts of Speech (PoS) features and 2 distinguishable features (D) as follows. The 45 PoS features were produced using the Stanford Part-of-Speech tagger (http://nlp.stanford.edu/software/tagger.shtml) (Toutanova et al., 2003). We normalized each one of the PoS features by the number of tokens of the post.

The D features were created from the labeled corpus (C1) as follows. Separately, for each sub-corpus (males, females) we activated the following process: The frequency of each unigram (word) in the corpus is counted (including stopwords). Only the unigrams with a frequency of at least 100 occurrences in the corpus are selected. For each sublist (males, females) only unigrams that appear in one sublist at least 1.2 times more than in the second sublist are selected as distinguishable unigrams. These distinguishable unigrams are sorted in non-ascending order, separately for each sublist. In this way, we obtained two long lists each of which contained several thousand words. For instance, for ratio factors of 2 or above (2+), the list of male unigrams consisted of 3401 words, and the list of female unigrams consisted of 2114 words. Table 1 introduces a few examples from these two lists generated from C1.

Table 1: Examples of distinguishable unigrams from the two lists with ratio factor of 2+.

Distinguishable males' unigrams	Distinguishable females' unigrams
scientists, criticism, war, bush, US, UN, Israel, terrorists, Linux, PC, Microsoft, networks, xbox, NBA, eBay, Greece, Sony, URL, beta, java, explorer, servers, government, businesses, SQL, rural, economics, constitutional, tactics, stance, versus, presidents, samurai, IP, mankind, enterprise, batman, FBI, civilians, GOP, veterans, CBS, NFL, Firefox, XML, electoral, IBM, .NET, NBC, commands	mom, baby, cuz, wedding, boyfriend, laundry, ugh, *sigh*, pregnant, jeans, yummy, butter, OMG, skirt, dishes, freaked, kisses, knitting, kitty, purse, makeup, sweater, outfit, bf, recipe, heels, photographer, dresses, massage, dork, oven, girly, nursing, babysitting, scarf, bake, flirt, comfy, feminist, LMAO, crushes, tanning, lipstick, brownies, muffin, witches, broccoli

We did not use these thousands of words as classical features, where each unigram is a feature. However, we applied and used only two features for each post. These two features represent the relative frequencies of the distinguishable males' unigrams / females' unigrams, i.e., the number of distinguishable males' unigrams / females' unigrams in a post normalized with respect to the total number of words in a post. We call these features: males' frequency and females' frequency, respectively.

5 THE CLUSTERING MODEL

Blog post clustering presents challenges due to the large number of potential features available, their dependencies, and the large number of training posts. Appropriate corpus construction, feature selection and text clustering are critical to the success of the clustering tasks.

We used two free available corpora. The first corpus (we call it C1) contains over 71,000 blogs including 681,288 blog posts with more than 140 million words (Schler et al., 2006). These blog posts were downloaded from Blogger.com one day in August 2004. Of all the bloggers: 37,324 are males and 34,169 are females. The C1 corpus is labeled. That is to say, for each post we know the gender of its author. Schler et al. (2006) performed gender classification tasks. They used 502 stylistic features (POS tags, blog words and hyperlinks) and 1000 unigrams with the highest information gain in the training set. They obtained an accuracy of 80.1%.

The second corpus (we call it C2) was published and distributed to the public by ICWSM (http://www.icwsm.org/2012/home/media/) (International AAAI Conference on Web and Social Media) in 2012. The C2's posts were downloaded from various blog web-sites, e.g., Blogger.com, WordPress, and LiveJournal. The C2 corpus is unlabeled. That is to say, for each post we do not know the gender of its author.

In view of the fact that the computers that were available to us were "modest" PCs and the run time was very long, it was decided to significantly "cut" the corpora and to base them only on posts (one for each selected blogger). Moreover, C2 included many noisy posts relatively to our clustering task, e.g., posts that were not written in English, posts containing sequence(s) of characters and symbols that are lack of context, and posts containing commercial ads.

C1 was filtered by us as follows. Firstly, we downloaded the 19,320 blogs that were chosen by Schler et al. (2006) for their classification task. For each one of these blogs, we selected only the first post. We received a new corpus including 19,320 posts. The filtered corpus was called C1B. Concerning C2, due to the relatively high percentage of noise data, we had to perform a long and slow manual filtering process. As a result, we received a relatively small-sized corpus containing 1,001 posts. This corpus is called C2B. Then, we cleaned each corpus by deletion of redundant spaces and normalization of words (e.g., transformation of letters from uppercase to lowercase and deletion of punctuation marks, e.g., ',', ';', '.', '!', '?', ''', and '"').

5.1 Selected Clustering Method

There is a wide range of clustering methods. Clustering methods can be divided into non-hierarchical methods such as K-means (Steinhaus, 1956) and K-means improved variants such as (Dhillon et al., 2002; Kanungo et al., 2002) and Expectation Maximization (EM) (Dempster et. al, 1977) and its improved variants such as (Bradley et al. 1998; Frenkel and Feder, 1999), and hierarchical methods such as hierarchical clustering (Johnson, 1967). Comprehensive surveys of various clustering methods and/or applications are for example: Jain et al. (1991), Zheng et al. (2006), and Aggarwal and Zhai (2012).

There is no agreed definition about what is a "good" or "correct" clustering method or as it was written by Estivill-Castro (2002) "clustering is in the

eye of the beholder". We decided to apply a commonly used cluster algorithm, which is called the expectation–maximization (EM) algorithm (Dempster et. al, 1977). This method uses an iterative computation of maximum-likelihood estimates. Each algorithm's iteration consists of an expectation step followed by a maximization step. This method assumes that the desired clusters have a normal distribution. The EM method with its default parameters has been applied using the Waikato Environment for Knowledge Analysis (WEKA) platform (Witten and Frank, 2005; Hall et al., 2009).

5.2 Feature Filtering

We decided to use the correlation feature selection (CFS) filtering method (Hall, 1999; Senliol et al., 2008) with its default parameters. CFS is the default filtering method implemented in WEKA. "CFS (Correlation-based Feature Selection) assumes that useful feature subsets contain features that are predictive of the class but uncorrelated with one another. CFS computes a heuristic measure of the "merit" of a feature subset from pair-wise feature correlations and a formula adapted from test theory. Heuristic search is used to traverse the space of feature subsets in reasonable time; the subset with the highest merit found during the search is reported." (Hall, 1999, p. 74).

5.3 Measurement of Results

For each experiment, we randomly choose 100 posts and for each post we checked whether the clustering choice was correct. The check was done by a mother tongue English speaker who read every word in the post and decided whether the author was a man or a woman. The accuracy rate for each experiment was defined according to the accuracy rate for the sample of 100 posts.

The number of clusters that is created by the EM method does not always match the number of groups that we intend to cluster. Often the number of the resulting clusters is higher. For instance, assuming clustering by gender, we can get five clusters while we expect only to two (males and females). When we look at the visual presentation of the resulting clusters, two clusters might be located on one side of the graph, the area that represents high values of feminine characteristics, while the other area contains the other three clusters representing high values of masculine characteristics. In this case, the two first clusters are defined as clusters of females, and the rest as clusters of males.

6 EXPERIMENTAL RESULTS

In this section, we present and analyze the clustering experiments of the personal blog posts included in the C2B corpus by gender using the EM clustering method, the CFS filtering method, and the PoS features for the C1B corpus and the D features extracted from C1B, as explained before.

In experiment # 1, we apply the CFS filtering method on the PoS features, and we get only 2 distinguishable features: DT (determiners) and PRP (pronouns). While DT consists of articles and is, therefore, more associated with males, PRP represents pronouns and is, therefore, more associated with females (Schwartz et al., 2013). The clustering accuracy result was only 59.78%. In experiment # 2, we did not apply the CFS filtering method. We applied EM on 4 features: DT, PRP and the two D features. The clustering accuracy result was slightly higher 60.4%. In experiment # 3, we applied the CFS filtering method on the PoS and the D features and we get only 12 distinguishable features: 11 PoS features and the females' frequency feature. The clustering accuracy result was 67.33%.
Table 2 and Figure 1 present 10 additional experiments where the accuracy of the gender clustering is presented as a function of either one D feature (males' frequency) or of the two D features. The results using only the females' frequency feature are not presented because they were significantly lower. A possible explanation to these findings is that the list of the males' distinguishable unigrams is much longer than the list of the females' distinguishable unigrams, as can be seen for ratio factor of at least 2, where the list of male unigrams consisted of 3,401 words while the list of female unigrams consisted of only 2,114 words. This could be because females write in a more consistent and straightforward way than males (Schwartz et al., 2013), which leads males to use a larger variety of words than females.

Table 2: Accuracy results according to the ratio factor.

Ratio factor		1.2+	1.4+	1.6+	1.8+	2+
Accuracy	Males' freq.	71.74	83.7	72.83	72.83	69.3
	Males' & females' freq.	70.65	80.43	70.65	70.65	68

A few interesting conclusions can be drawn from the gender clustering experiments: (1) the males' frequency feature was found as the best clustering feature, (2) among the various values tested for the

ratio factor for the males' frequency feature, the optimized accuracy result (83.7%) was obtained by a ratio factor of 1.4+. Increase or decrease in this ratio factor leads to lower results, and (3) using both the males' and the females' frequency features with the same values of the ratio factors lead to a curve similar to the curve that is based on the males' frequency feature only but with lower values.

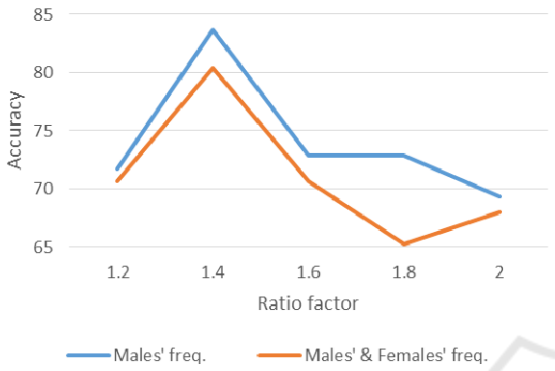

Figure 1: Accuracy of the gender clustering model as a function of the ratio factor of two types of features.

In our experiments, the CFS filtering method did not select the features that led to the best accuracy results. Strict manual selection and removal of ineffective features done by trial and error was much more successful.

As mentioned before, to the best of our knowledge, our research is the first to cluster blog posts by gender. Therefore, we cannot compare our method with state-of-the-art methods. A very partial comparison can be made with the gender classification study on the C1 corpus (Schler et al., 2006) achieving an accuracy of 80.1%. Our clustering method achieves a slightly better accuracy result (83.7%) indeed on another corpus (C2B). Furthermore, our result is quite promising because it is known that clustering is regarded as a difficult task than classification.

7 SUMMARY AND FUTURE WORK

In this paper, we propose a methodology for effective clustering of personal blog posts written in English by gender. To the best of our knowledge, we present two novelties: (1) this is the first study to cluster blog posts by gender, and (2) the use of distinguishable features that were extracted from a labeled corpus for the clustering of another similar corpus, which is not labeled.

We constructed two filtered corpora from two free available personal blog corpora. Each blog from the first corpus was already gender-labeled while the blogs of the second corpus were not labeled. We extracted from the labeled corpus distinguishable unigrams for both males and females. Then, we defined two features that represent the relative frequencies of the distinguishable males' unigrams and females' unigrams, (males' and females' frequency). The best distinguishable feature was found to be the males' frequency feature with a ratio factor of 1.4 relatively to females. This feature leads to an accuracy rate of 83.7% for the gender clustering task of a totally different blog corpus, which is unlabeled.

There is much room for future research in this domain. Possible directions are: (1) Applying additional feature sets for the clustering tasks, e.g., N-grams (N > 2), function words, orthographic features, quantitative features, and topographic features. These features can help with style-based clustering and might help detect differences in style that are characteristic of the gender of the author as well as with other tasks; (2) Conducting additional experiments using much larger unlabeled blog posts corpora and additional combinations of ratio factor and minimal frequencies of distinguishable N-grams; (3) Comparing between the proposed method and other methods such as (a) gender-based supervised classification, and (b) simple document clustering exploiting, e.g., textual similarity; and (4) Extending the experiments to other interesting clustering tasks such as clustering by age, personality type, political orientation, and regional origin.

REFERENCES

Aggarwal, C. C., Zhai, C., 2012. A survey of text clustering algorithms. *In mining text data (pp. 77-128). Springer US.*

Aldenderfer, M. S., Blashfield, R. K., 1984. Cluster Analysis Sage University Papers Series. Quantitative.

Bailey, Ken., 1994. *Numerical taxonomy and cluster analysis. Typologies and Taxonomies. p. 34.*

Banerjee, S., Ramanathan, K., Gupta, A., 2007. Clustering short texts using Wikipedia. *In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 787-788). ACM.*

Bradley, P. S., Fayyad, U., Reina, C., 1998. Scaling EM (expectation-maximization) clustering to large

databases (pp. 9-15). *Redmond: Technical Report MSR-TR-98-35, Microsoft Research.*

Brooks, C., Montanez, N., 2006. Improved annotation of the blogosphere via auto tagging and hierarchical clustering, in: *Proceedings of the WWW 2006, ACM, Edinburgh, UK, 625–632.*

Mary Bucholtz and Kira Hall. 2005. Identity and interaction: A sociocultural linguistic approach. Discourse studies, 7(4-5):585–614.

Burger, J. D., Henderson, J., Kim, G., Zarrella, G., 2011. Discriminating gender on Twitter. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1301-1309). Association for Computational Linguistics.*

Cucchiara, R. 1998. Genetic algorithms for clustering in machine vision. Machine Vision and Applications, 11(1), 1-6.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B 39 (1): 1–38.*

Dhillon, I. S., Guan, Y., Kogan, J., 2002. Iterative clustering of high dimensional text data augmented by local search. *In Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on (pp. 131-138). IEEE.*

Eckert, P., McConnell-Ginet. S., 2013. Language and gender. *Cambridge University Press.*

Eckert, P., 1997. Age as a sociolinguistic variable. The handbook of sociolinguistics. Blackwell Publishers.

Estivill-Castro, Vladimir. 2002. Why so many clustering algorithms — A Position Paper.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. 1996. Advances in knowledge discovery and data mining.

Frenkel, L., Feder, M., 1999. Recursive expectation-maximization (EM) algorithms for time-varying parameters with applications to multiple target tracking. *IEEE Transactions on Signal Processing, 47(2), 306-320.*

Filho, J. A. B. L., Pasti, R., de Castro, L. N., 2016. Gender Classification of Twitter Data Based on Textual Meta-Attributes Extraction. *In New Advances in Information Systems and Technologies (pp. 1025-1034). Springer International Publishing.*

Gao, J., and Lai, W. 2010. Formal concept analysis based clustering for blog network visualization. *In Advanced Data Mining and Applications (pp. 394-404). Springer Berlin Heidelberg.*

HaCohen-Kerner, Y., Margaliot, O., 2013. Various document clustering tasks using word lists. In Asia Information Retrieval Symposium (pp. 156-169). *Springer Berlin Heidelberg.*

HaCohen-Kerner, Y., Margaliot, O., 2014. Authorship attribution of responsa using clustering. Cybernetics and Systems, 45(6), 530-545.

Hall, M. A. 1999. Correlation-based feature selection for machine learning (Doctoral dissertation, The University of Waikato).

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten., 2009. The WEKA Data Mining Software: an Update. *ACM SIGKDD Explorations Newsletter, 11(1), pp.10-18.*

Jain, A. K., Murty, M. N., Flynn, P. J., 1991. Data Clustering: A Review. *ACM Computing Surveys 31, 3 (264–323).*

Johnson, S. C., 1967. Hierarchical clustering schemes. Psychometrika, 32(3), 241-254.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., Wu, A. Y., 2002. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence, 24(7), 881-892.*

Koppel, M., Argamon, S., Shimoni, A. R., 2002. Automatically categorizing written texts by author gender. Literary and Linguistic Computing, 17(4), 401-412.

Kuzar, T., Navrat, P. 2011. Slovak blog clustering enhanced by mining the web comments. *In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on, Vol. 3, 293-296. IEEE.*

Marwick, A. E. and Boyd D., 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. New Media Society, 13(1):114–133.

Mukherjee, A., Liu, B., 2010. Improving gender classification of blog authors. *In Proceedings of the 2010 conference on Empirical Methods in natural Language Processing (pp. 207-217). Association for Computational Linguistics.*

Ngan, M., and Grother, P. 2015. Face recognition vendor test (frvt) performance of automated gender classification algorithms. *In Technical Report NIST IR 8052. National Institute of Standards and Technology.*

Nguyen, D. P., Trieschnigg, R. B., Doğruöz, A. S., Gravel, R., Theune, M., Meder, T., de Jong, F. M. G. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. COLING, Association for Computational Linguistics.

Yingbo Miao, Vlado Keselj, and Evangelos Milios. Document Clustering using Character N-grams: A Comparative Evaluation with Term-based and Word-based Clustering. *In Proc. of the 14th ACM int. conference on Information and knowledge management, 357–358. 2005.*

Schler, J., Koppel, M., Argamon, S., Pennebaker, J. W., 2006. Effects of Age and Gender on Blogging. *In AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, Vol. 6, pp. 199-205. AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. Vol. 6. 2006.*

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... Ungar, L. H., 2013. Personality, gender, and age in the language of social media: *The open-vocabulary approach. PloS one, 8(9), e73791.*

Sharan, R., Shamir, R., 2000. CLICK: a clustering algorithm with applications to gene expression analysis. *In Proc Int Conf Intell Syst Mol Biol (Vol.8, No.307, p.16).*

Steinhaus, H., 1956. Sur la division des corp materiels en parties. Bull. Acad. Polon. Sci, 1(804), 801.

Tasoulis, D. K., Plagianakos, V. P., and Vrahatis, M. N. 2004. Unsupervised clustering of bioinformatics data. *In European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems, Eunite, pp. 47-53.*

Toutanova, K., Klein, D., Manning, C. D., Singer, Y.: Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network. *In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Vol. 1, NAACL'03, Association for Computational Linguistics, 173–180. 2003.*

Tryon, R. C., 1939. Cluster Analysis: Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality. Edwards Brothers.

Witten, I. H., E. Frank., 2005. Data Mining: Practical Machine Learning Tools Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems). San Mateo, CA: Morgan Kaufmann.

Yan, X., and Yan, L., 2006. Gender Classification of Weblog Authors. *In AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs (pp. 228-230).*

Zheng, Y., Cheng, X., Huang, R., Man, Y., 2006. A comparative study on text clustering methods. *In International Conference on Advanced Data Mining and Applications (pp. 644-651). Springer Berlin Heidelberg.*