Authorship Attribution in Health Forums

Victoria Bobicev
Technical University of Moldova
Chisinau, Moldova
vika@rol.md

Marina Sokolova
CHEO Research Institute, Ottawa
University of Ottawa, Ontario, Canada
sokolova@uottawa.ca

Khaled El Emam
CHEO Research Institute, Ottawa
University of Ottawa, Ontario, Canada
kelemam@uottawa.ca

Stan Matwin
Dalhousie University, Halifax, Canada
Polish Academy of Sciences, Warsaw, Poland
stan@cs.dal.ca

Abstract

The emergence of social media (networks, blogs, web forums) has given people numerous opportunities to share their personal stories, including details of their health. Although users mostly post under assumed nicknames, state-of-the-art text analysis techniques can combine texts from different media and use that linkage to identify private details of an individual's health. In this study we aim to empirically examine the accuracy of identifying authors of on-line posts on a medical forum.[1] Our results show a high accuracy of the authorship attribution, especially when text is represented by the orthographical features.

1 Introduction

Emergence of social media (networks, blogs, web forums) has given people numerous opportunities to share their personal stories, including details of their health (e.g., disease diagnosis, symptoms, treatment) (Velden and Emam, 2012; Bobicev et at, 2012):

- The transfer went well - my RE did it himself which was comforting. 2 embies (grade 1 but slow in development) so I am not holding my breath for a positive.
- I've had 7 IUI and one ivf all cancelled due to not ovulating. I am a poor responder. What bothers me the most is never getting to the point of actually going thru the procedure.[2]

Sharing personal health information (PHI) is a behavior that can be seen in 80% of Internet users, or in 59% of all adults, who reported searching for health information (Fox, 2011).

Although users mostly post under assumed nicknames, state-of-the-art text analysis techniques can combine texts from different forums and then use that linkage to identify private details of an individual's health. Aggregating and mining posts from five forums, Li et al. (2011) identified the user's full name, date of birth, spouse's name, home address, home phone number, cell phone number, email, occupation and the lab test results. The latter are highly indicative of the suspected disease, and hence, of the health conditions of the said individual.

In order to gauge how best to protect internet user anonymity, we first wanted to know the ability of Text Mining techniques in authorship attribution on medical forums, i.e. the task of identification of an author among other authors posting on the same forum. The attribution is based on comparison of a new text to texts previously written by known authors.

We obtained the empirical evidence on the posts from an on-line community of IVF (In Vitro Fertilization) patients. We achieved a highly accurate authorship attribution: up to 90% when the text is represented by the orthographical features.

[1] This work had been done when the first author was a visiting professor at CHEO Research Institute.

[2] The messages have an original spelling and punctuation.

Proceedings of Recent Advances in Natural Language Processing, pages 74–82,
Hissar, Bulgaria, 7-13 September 2013.

2 Related works

Authorship attribution has been intensively investigated by Computational Linguistics. Starting 2007, an annual competition on author attribution has been organized in conjunction with CLEF. [3]

Accuracy of the authorship attribution depends on features extracted from the analyzed text. Vocabulary features used in various research are word length (Brinegar, 1963), sentence length (Morton, 1965), vocabulary richness (Tweedie and Baayen, 1998), word n-gram frequencies (Hoover, 2003), errors and idiosyncrasies (Koppel and Schler, 2003), synonyms and semantic dependencies (Afroz et al., 2012).

A few studies used syntactic features, e.g. parts of speech and part of speech sequences (Zhao and Zobel, 2007), chunks of text (Stamatatos et al, 2001), syntactic dependencies of words (Gerritsen, 2003), and syntactic structures (Hirst and Feiguina, 2007).

The use of orthographical features in the attribution task was studied in Abbasi and Chen (2008). The features included characters, characters bigrams and trigrams, punctuation and special characters, as well as common vocabulary features. 88-96% accuracy was achieved on several data sets including e-bay comments, Java forum, email and chat corpora. Narayanan et al. (2012) adapted this feature set in the author classification of 100,000 blogs where the average length of each blog was 7500 words. The paper's authors correctly identified an anonymous author in >20% of cases; in approximately 35% of cases the correct author was one of the top 20 guesses. At the same time, Koppel (2009) had shown that 1000 character trigrams with highest information gain helped SVM to obtain 80-86% in attribution accuracy on literature corpus, email and blog corpora.

With the emergence of user-written Web content, authorship analysis is often done on online messages (Zheng et al., 2006; Narayanan et al., 2012). Large numbers of candidate authors, small volumes of training and test texts, and short length of messages makes the online authorship analysis exceptionally challenging (Juola, 2006; Koppel, 2009; Luyckx and Daelemans, 2008; Madigan et al., 2005; Stamatatos, 2009).
In Koppel et al. (2006), 10,000 blogs were used in the task of author attribution. The test data was built from 500-word snippets, one for each author. 20-34% of texts were classified with average accuracy of 80%; the rest of texts were considered unknown. In Koppel et al. (2011), on the same dataset, a 500-word snippet was attributed to one of 1,000 authors with *Coverage* = 42.2% and *Precision* = 93.2%. Consequently, the remaining 57.8% of snippets were considered unknown.

None of these cited works, however, considered authorship analysis of messages posted on medical forums or other online venues that are dedicated to discussions of personal health information.

3 The Forum Data

We focused on the authorship attribution on medical forums where the authors may post sensitive PHI, e.g., problems with conception. In particular, we worked with data from IVF.ca, an infertility on-line community created by prospective, existing and past IVF (In Vitro Fertilization) patients. The IVF.ca website includes forums: *Cycle Friends*, *Expert Panel*, *Trying to Conceive*, *Socialize*, *In Our Hearts*, *Pregnancy*, *Parenting*, and *Administration*.

The forums listed above consist of several sub-forums, e.g., the *Cycle Friends* forum consists of *Introductions*, *IVF/FET/IUI Cycle Buddies*, *IVF Ages 35+ and other*. Every sub-forum contains of a number of topics initiated by a forum participant, e.g. the "*IVF Ages 35+"* subforum contains 506 topics such as "*40+ and chances of success*", "*Over 40 and pregnant or trying to be*", etc. Depending on the topic itself and the amount of interest among participants, different numbers of posts are associated with each topic. For example, "*40+ and chances of success*" has four posts and "*Over 40 and pregnant or trying to be*" has 1136 posts.

Note that differentiation between the authors of posts is easier when the authors exhibit contrasting writing styles. The style dissimilarity usually comes with diversity among the author population and the topics they write about (Koppel et al., 2009).

We, on the other hand, worked with the forum posts that lack such diversity. Hence, the texts are more complex in differentiation between the authors. Specifically:

a) the posts have a unified content (i.e., all posts are about infertility treatment);

b) the same gender of authors (i.e., participants are overwhelmingly women);

[3] http://pan.webis.de

c) a small age range (most authors are 35-40 years old);

d) the same geographic location (most are Canadians and a few USA);

e) the same time of posting (2008 - 2012).

We intended to use posts as analysis units, i.e. our goal was to identify the author of each post individually. We assumed that the length of the texts written by an author would be sufficient for a meaningful analysis and that we needed a substantial number of posts per author. Two sub-forums *IVF Ages 35+* and *Cycle Buddies* satisfied our criteria better than other sub-forums.

We grouped posts by the authors to estimate the amount of text every author wrote and sorted these estimates according to the number of posts written by each author in descending order. Only a small number of authors had many posts. The post-per-author distribution for the first 100 of the most prolific authors in both forums is presented on Figure 1.

Only the first 30 authors in the *Age 35+* sub-forum had more than 100 posts; in the *Cycle Buddies* sub-forum situation was a little better, as almost all the 100 first authors had more than 100 posts. However, many posts contained citations of other authors and only short replies and we had to remove such posts from further studies.

The average length of posts was also important as shorter messages were harder to identify. The average length of posts in the *Ages 35+* sub-forum was about 750 characters (approx. 150 words) and in the *Cycle Buddies* subforum - about 600 characters or approx. 100 words. The larger number of posts in this sub-forum allowed us to remove the shortest posts and posts with citations.

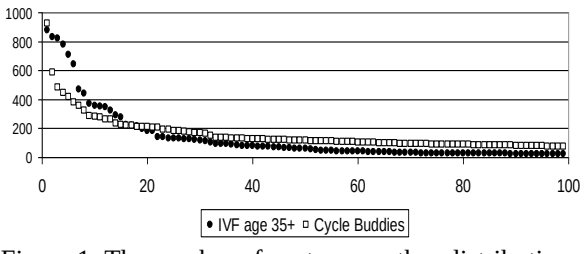

Figure 1: The number of posts per author distribution for the first 100 authors

For the empirical experiments, we harvested 18685 messages from the most prolific 30 authors from every forum, i.e. 60 authors in total, and selected 100 messages per an author for future analysis. We worked exclusively with the message contents. No author metadata was used in the file analysis.

It should be noted that most of the selected authors posted in many different topics and we collected posts without exclusion of any topics. Thus author classification had no influence of topic differences. Figure 2 presents the numbers of topics in which the 30 authors whom we selected for the experiments from *Age 35+* sub-forum posted.

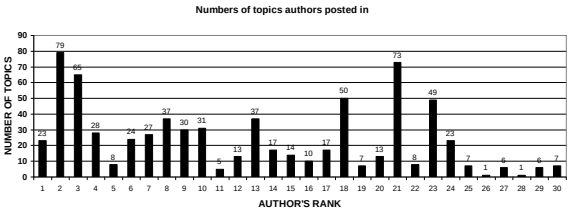

Figure 2: The number of topics the authors posted in for the first 30 authors of Age 35+ sub-forum.

4 Stylistic Features and Authorship Attribution

The authorship attribution task traditionally relies on

- a statistical analysis of the author's vocabulary, e.g., the number of distinct words, occurrences of words, identification of most frequent words and phrases;
- the analysis of the composition style, e.g., position of words in sentences, type and length of sentences, paragraph formation (Oakes, 2005).

Provided there was enough data for quantitative analysis, the results of these analyses were able to accurately attribute authorship. The requirement usually implied a minimum of five occurrences of a feature.

Texts gathered from the web forums were usually short. In our data, an average post had one or two paragraphs and 50-250 words. A small number of occurrences of words determined the type of features we could use in our authorship attribution task. For example, even after combining all the posts of the same author in one document, we still could not meaningfully use the composition-style features for authorship attribution.

Choosing from the vocabulary features, we could use the most frequent words but not phrases. The vocabulary statistics would not be reliable as well, due to a small corpus size for each author.

At the same time, we had sufficient quantities of the orthographical features per author to use them in the authorship attribution. These features included alphabetic and non-alphabetic characters, capitalization, and punctuation. Currently, the orthographical features were often used to analyze short text messages, e.g. tweets. Common tasks included named entity recognition (Ritter et al., 2011) and text normalization (Han and Baldwin, 2011). The features were used in the authorship attribution through language modeling (Peng et al., 2003) and machine learning (Koppel and Schler, 2003).

4.1 Vocabulary features

Our initial word set was the same for both subforums. The set of the most frequent words consisted of 50 words that sometimes are referred to as 'stop' or 'short' words (me, of, get, have). Such words are often removed in text classification. However, they played an important role in the authorship attribution task (Zhao and Zobel, 2005). The rest of the used 3796 words (egg, wish), were salient words with frequency > 3 in the frequency dictionary for the joint sub-forum data.

To reduce redundancy of the features, we removed words that did not discriminate between the authors. The resulting feature sets considerably varied.

Table 1 shows the numbers of the vocabulary features for both sub-forums. We introduced the features' ID for the further reference.

Features	ID	Cycle Buddies	Age 35+
Frequent words	I	50	50
Salient words	II	3583	3788
All words	III	3633	3838

Table 1: The vocabulary features for the Cycle Buddies and Age 35+ subforums.

Features	ID	Cycle Buddies/Age 35+
Lower case letters	IV	26
Capital and low letters	V	52
Punctuation	VI	24
Numbers and punctuation	VII	34
All characters	VIII	86

Table 2: The orthographical features for the data.

4.2 Orthographical features

We used standard orthographical features, such as lower-case letters (a - z), capitalization (C, c), punctuation (;,!), etc. Table 2 reports the categories of the features and the number of features in each category. Feature numbers were the same for both subforums. Again, we introduced the features' ID for further reference in machine learning experiments.

4.3 Combined features

We used two feature sets that were combined from the vocabulary and the orthographical features.

The first set was an unaltered combination of all the features without useless features (i.e., features that did not discriminate among classes were removed). Another set was an outcome of the BestFirst selection algorithm; this set included punctuation (?, ., !), letters (e, n) and words (ladies, thanks, two, transfer).

Features	ID	#
Useless features removed	IX	3719
BestFirst selected features	X	73

Table 3: Combined features for the Cycle Buddies data.

Tables 3 and 4 list the number of features for the *Cycle Buddies* and the *Age 35+* sub-forums.

Features	ID	#
Useless features removed	IX	3924
BestFirst selected features	X	75

Table 4: Combined features for the Age 35+ data.

5 Machine Learning Experiments

In our previous work in classification of short texts (Bobicev et al., 2012), Naïve Bayes had been shown as highly accurate when compared with other ML algorithms. Due to NB's high efficiency we opted to apply it as well as KNN, another highly efficient algorithm. This task was solved as a multi-class classification problem, where one class represented one author. There were 30 authors in each subforum, hence that data sets were categorized into 30 classes.

We assessed the learning methods by computing multi-class *Precision (Pr), Recall (R), F-score (F)* and *Accuracy (Acc):*

$Precision = \sum_{i=1}^{n} \frac{tp_i}{tp_i + fp_i}$ is the ratio of texts belonging to categories $c_1,...,c_n$ to all texts classified to these categories.

$Recall = \sum_{i=1}^{n} \frac{tp_i}{tp_i + fn_i}$ is the percentage of texts belonging to categories $c_1,...,c_n$ that are indeed classified into these categories.

We use the balanced *F-score* which is the harmonic mean of *Precision* (P) and *Recall* (R):

$$F\text{-}score = 2PrR / (Pr + R)$$

$Accuracy = \frac{1}{n} \sum_{i=1}^{n} \frac{tp_i + tn_i}{tp_i + fn_i + tn_i + fp_i}$ is the average *Accuracy* obtained on all the categories.

In these formulae, tp_i is the number of texts classified into the category c_i that indeed belong to c_i, fp_i is the number of texts classified into c_i that do not belong to c_i, fn_i is the number of texts that indeed belong to c_i but were not classified into it, tn_i is the number of texts that do not belong to c_i and were not classified into it.

Data	Pr	R	F	Acc (%)
Cycle Buddies	0.002	0.043	0.040	4.33
Age 35+	0.001	0.034	0.020	3.37

Table 5: Baseline classification results.

Features	Naïve Bayes			
	Pr	R	F	Acc (%)
I	0.385	0.386	0.380	38.64
II	**0.714**	**0.635**	**0.648**	**63. 55**
III	***0.683***	***0.580***	***0.594***	***57.98***
IV	0.212	0.225	0.213	22.45
V	0.374	0.360	0.359	35.96
VI	0.379	0.365	0.354	36.45
VII	0.403	0.370	0.365	36.97
VIII	0.564	0.541	0.533	54.11
IX	0.648	0.524	0.520	52.44
X	***0.625***	***0.557***	***0.544***	***55.73***

Table 6: NB classification of the Cycle Buddies data.

For the baseline performance evaluation, we chose classification of all authors into the largest class. Table 5 presents the baseline classification results for the subforums.

We applied 10-fold cross-validation for the best classifier selection. Each post was used as an independent element. Thus, in each run of 10-fold cross-validation for each author 90 posts were used for training and 10 posts functioned as test items. The author was identified for each of them; hence we had 30 classes with 90 posts for training and 300 test posts. Tables 6 and 7 report the best classification results of both algorithms on each feature set for the Buddies subforum. Tables 8 and 9 report the best classification results for the both algorithms on the Age 35+ subforum. We put the top result for each classifier in **this font**. We mark the second and the third best results with ***this font.***

Features	K-Nearest Neighbor			
	Pr	R	F	Acc (%)
I	0.266	0.218	0.223	21.85
II	0.374	0.125	0.131	12.50
III	0.350	0.130	0.134	12.96
IV	0.185	0.160	0.159	16.04
V	0.293	0.259	0.261	25.89
VI	***0.375***	***0.352***	***0.354***	***35.15***
VII	0.355	0.322	0.327	32.24
VIII	**0.413**	**0.381**	**0.382**	**38.07**
IX	0.360	0.137	0.140	13.65
X	***0.420***	***0.364***	***0.372***	***36.36***

Table 7: KNN classification of the Cycle Buddies data.

Features	Naïve Bayes			
	Pr	R	F	Acc (%)
I	0.399	0.411	0.400	41.08
II	**0.770**	**0.681**	**0.696**	**68.08**
III	***0.730***	***0.622***	***0.639***	***62.19***
IV	0.215	0.233	0.216	23.30
V	0.331	0.342	0.330	34.24
VI	0.382	0.359	0.351	35.86
VII	0.387	0.372	0.364	37.17
VIII	0.544	0.539	0.527	53.87
IX	***0.680***	***0.560***	***0.561***	***55.99***
X	0.611	0.549	0.532	54.95

Table 8: NB classification of the Age 35+ data.

The presented results show that NB performs better than KNN on both forums. Moreover, this holds true for all the 10 feature sets in the forums.

From the combined features only the set X (i.e., BestFirst selected features) provided rea-

sonably good results. The set IX (i.e., all features but useless) did not provide a reliable classification.

Features	K-Nearest Neighbor			
	Pr	**R**	**F**	**Acc (%)**
I	0.317	0.282	0.279	28.25
II	0.419	0.140	0.127	14.04
III	0.375	0.144	0.129	14.38
IV	0.197	0.185	0.180	18.52
V	0.310	0.285	0.280	28.49
VI	***0.323***	***0.304***	***0.298***	***30.44***
VII	0.298	0.279	0.273	27.90
VIII	***0.400***	***0.363***	***0.359***	***36.33***
IX	0.431	0.145	0.132	14.55
X	**0.459**	**0.423**	**0.425**	**42.26**

Table 9: KNN classification of the Age 35+ data.

The most striking difference in the classifier performance is found on Features II, i.e. low and capital letters. On this feature set, NB achieves its best performance on both forums (F = 0.648 for the *Cycle Buddies*, F = 0.696 for the *Age 35+*), while KNN has its worst performance on the forums (F = 0.131 for the *Cycle Buddies*, F = 0.127 for the *Age 35+*).

6 Model-based Authorship Attribution

In this part of our work,we the language model-based attribution. We used Prediction by Partial Matching (PPM statistical model) for authorship classification. Prediction by Partial Matching (PPM) is an adaptive, finite-context method for text compression (Cleary, Witten, 1984).

An example of the general method of context probability interpolation is the probability of character '***l***' in the context of the word '***medical***' calculated as a sum of conditional probabilities of this character in dependence of different context length up to the limited maximal length in this particular case equal to 5:

$$P_{blended}('\mathbf{l}') = \lambda_5 \cdot P('\mathbf{l}' \mid '\mathbf{edica}') + \lambda_4 \cdot P('\mathbf{l}' \mid '\mathbf{dica}') + \lambda_3 \cdot P('\mathbf{l}' \mid '\mathbf{ica}') + \lambda_2 \cdot P('\mathbf{l}' \mid '\mathbf{ca}') + + \lambda_1 \cdot P('\mathbf{l}' \mid '\mathbf{a}') + \lambda_0 \cdot P('\mathbf{l}')$$

where λ_i (i = 1…5) are normalization coefficients; some of them can be equal to zero and $\sum_{i=1}^{5} \lambda_i = 1$, where 5 is the maximal length of the context.

Bratko and Filipic (2005) used letter-based PPM models for spam detection. In this task there existed two classes only: spam and legitimate email (ham). The created models showed strong performance in Text Retrieval Conference competition, indicating that data-compression models are well suited to the spam filtering problem.

Teahan et al. (2000) used a PPM-based text model and minimum cross-entropy as a text classifier for various tasks including the author attribution for the well known Federalist Papers.

Bobicev and Sokolova (2008) applied the PPM algorithm for text categorization. They used character-based and word-based PPM. The character-based PPM outperformed the word-based PPM.

In the current work we applied PPM to the orthographical features described in Section 4.2.

6.1 Classification Experiments

As in previous experiments, we used 10-fold cross-validation for the best model selection.

Tables 10 and 11 present results for the both sub-forums. We put the top results for each forum in **this font**. We mark the second and the third best results with ***this font.***

Features	Pr	R	F	Acc (%)
IV	***0.851***	***0.822***	***0.836***	***82.2***
V	***0.882***	***0.857***	***0.869***	***85.7***
VI	0.400	0.363	0.380	36.3
VII	0.391	0.387	0.389	38.7
VIII	**0.911**	**0.893**	**0.902**	**89.4**

Table 10: Classification of the Cycle Buddies data.

Features	Pr	R	F	Acc (%)
IV	***0.761***	***0.743***	***0.752***	***74.3***
V	***0.797***	***0.777***	***0. 787***	***77.7***
VI	0.331	0.325	0. 328	32.5
VII	0.368	0.357	0.362	35.7
VIII	**0.836**	**0.817**	**0.826**	**81.7**

Table 11: Classification of the Age 35+ data.

The empirical results show that model-based classification of authors significantly outperforms probability-based and prototype-based classification when applied to both the letter and all the characters features. All three algorithms

achieve approximately the same accuracy when applied to punctuation and number features.

7 Discussion

We have shown empirically that stylistic features can help to identify an author among a large group of authors. Solving 30-class classification problems for two subforums, we constantly outperformed the baseline classification. Application of Naïve Bayes on the vocabulary features gave the best overall results for authorship attribution on the both subforums.

In general, Naïve Bayes performed better on the vocabulary features than on the orthographical ones; the reverse was true for KNN. However, Naïve Bayes outperformed K-Nearest Neighbor on the orthographical features as well.

Comparison of the best performance of the two algorithms showed that a probabilistic algorithm significantly outperforms a prototype algorithm in the authorship attribution on the medical subforum data.

The most impressive *Accuracy* and *F-score* gains were obtained by application of the model-based PPM on the letter and all-character features. The algorithm outperformed NB and KNN on both the forums. However, the specific PPM methodology of feature use makes much more difficult the comparison of the influence of specific text features on the author attribution task performance.

It should be noted that we obtained these results using internet forum posts and the length of these posts varied considerably. There were posts consisting of two or three words, e.g. "good luck!". We were able to identify the authors of the longer texts with an accuracy of 90%.

We also noticed that longer posts often contained important and sensitive information about person's health. If accessed and generalized from several posts, this extensive health information can be potentially harmful for the author. Personal and health information can be too extensive if, for example, it reveals the location, the diagnosis, and contains a possibility to identify the name. For example, in one post a patient says in what hospital she has a treatment, i.e. identifying the location. In another posts she specifies the treatment (this can also hint on the costs, hence, the income/money range) and she refers to a friend/relative giving their names. Or a patient complains about a specific condition (e.g., being overweight), telling others in what area she lives in and to what specialist (e.g., obesity doctor) she goes for treatment. These facts can be combined to create an accurate estimation of the poster's identity. Both listed scenarios present real cases that we've found in the data.

8 Conclusions

In this study we empirically examined the accuracy of identifying authors of online posts on a medical forum. Given that individuals may be reluctant to share personal health information on online forums, they may choose to post anonymously. The ability to determine the identity of anonymous posts by analyzing the specific features of the text raises questions about users posting anonymously as a method to control what is known publicly about them.

We have shown that the application of learning methods, especially NB and PPM, makes an automated identification of the author of an online post possible. Our method was able to correctly attribute authors with high confidence.

The focus of this work has been to show that the vocabulary and orthographical features can help to identify authors with a degree of high accuracy. Our experiments show that the authorship attribution based on orthographical features can be more effective that the authorship attribution based on the vocabulary features. We hypothesize that the use of orthographical features reflects on the author's personality. For example, in emotionally rich posts, the authors excessively use punctuation to emphasize their sentiments (e.g., question and exclamation marks, emoticons); those features are specific for each author.

To reduce the risk of a possible identification, we can suggest the author to change his or her habits of capitalization and the use of punctuation marks, as well as the use of emoticons.

These results are novel for the forum analysis, as the usual text analysis methods are based on semantics and analyze the use of words, phrases and other text segments.

The main implication of our results is that managers of online properties that encourage user input should also alert their users about the strength of anonymity. They should also caution users from posting sensitive information anonymously.

Acknowledgments

This work has been funded by the NSERC Discovery and Strategic Projects Research Grants. The authors thank Brian Dewar for his help with editing the manuscript.

References

Abbasi A. and Chen H. 2008. *Writeprints: a stylometric approach to identity-level identification and similarity detection in cyberspace.* ACM Transactions on Information Systems 2008; 26(2):1-29.

Afroz S., Brennan M., Greenstadt R. 2012. *Detecting Hoaxes, Frauds, and Deception in Writing Style Online.* IEEE Symposium on Security and Privacy 2012: 461-475.

Argamon S., Koppel M., Pennebaker J. and Schler J. 2008. *Automatically Profiling the Author of an Anonymous Text*. Communications of the ACM Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503-512.

Bobicev V. and Sokolova M. 2008. *An effective and robust method for short text classification.* Proceedings of the 23rd national conference on Artificial intelligence - Volume 3, 2008, pp. 1444–1445.

Victoria Bobicev, Marina Sokolova, Yasser Jafer, David Schramm. 2012. *Learning Sentiments from Tweets with Personal Health Information*. Canadian Conference on Artificial Intelligence 2012: 37-48.

Bratko A. and Filipic B. 2005. *Spam Filtering Using Compression Models*. Department of Intelligent Systems, Jozef Stefan Institute, Ljubljana, Slovenia, IJS-DP-9227.

Michael Brennan and Rachel Greenstadt. 2009. *Practical Attacks Against Authorship Recognition Techniques.* Proceedings of the Twenty-First Conference on Innovative Applications of Artificial Intelligence (IAAI).

Brinegar C. S. 1963. *Mark Twain and the Quintus Curtius Snodgrass Letters: A Statistical Test of Authorship.* Journal of the American Statistical Association 58, pp. 85–96.

Cleary J. and Witten I. 1984. *Data Compression Using Adaptive Coding and Partial String Matching*. IEEE Transactions on Communications, vol. 32, no. 4, pp. 396 – 402.

Bo Han and Timothy Baldwin. 2011. *Lexical normalisation of short text messages: Makn sens a #twitter*. ACL 2011.

Fox S. 2011. *The Social Life of Health Information, 2011*. the survey report http://www.pewinternet.org/Reports/2011/Social-Life-of-Health-Info.aspx.

Gerritsen C. M. 2003. *Authorship Attribution Using Lexical Attraction.* M.S. Dissertation, Massachusetts Institute of Technology.

Hirst G. and Feiguina O. 2007. *Bigrams of syntactic labels for authorship discrimination of short texts*. Literary and Linguistic Computing, 22(4), pp. 405-417.

Hoover D. L. 2003. *Frequent Collocations and Authorial Style*. Literary and Linguistic Computing 18: 261–286.

Patrick Juola. 2006. *Authorship Attribution*. Foundations and Trends® in Information Retrieval: Vol. 1: No 3, pp 233-334.

Koppel M. and Schler J. 2003. *Exploiting Stylistic Idiosyncrasies for Authorship Attribution*. Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, pp. 69-72.

Koppel M. and Schler J. 2004. *Authorship verification as a one-class classification problem.* Proceedings of the 21st International Conference on Machine Learning.

M. Koppel, J. Schler, S. Argamon, and E. Messeri. 2006. *Authorship attribution with thousands of candidate authors*. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, USA, 2006, pp. 659–660.

Koppel M., Schler J. and Argamon S. 2009. *Computational methods in authorship attribution*. JASIST 60 (1): 9–26.

M. Koppel, J. Schler, and S. Argamon. 2011. *Authorship attribution in the wild*. Lang Resources & Evaluation, vol. 45, no. 1, pp. 83–94.

Kukushkina O.V., Polikarpov A.A., and Khmelev D.V. 2001. *Using literal and grammatical statistics for authorship attribution.* Problems of Information Transmission, 37(2), 172-184.

F. Li, X. Zou, P. Liu, and J. Y. Chen. 2011. *New threats to health data privacy*. BMC Bioinformatics, vol. 12 Suppl 12, p. S7.

Kim Luyckx and Walter Daelemans. 2008. *Authorship Attribution and Verification with Many Authors and Limited Data.* Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008), 513-520.

D. Madigan, A. Genkin, D. D. Lewis, S. Argamon, D. Fradkin, and L. Ye. 2005. *Author Identification on the Large Scale*. Joint Meeting of the Interface and Classification Society of North America.

Andrew McDonald, Sadia Afroz, Aylin Caliskan, Ariel Stolerman and Rachel Greenstadt. 2012. *Use Fewer Instances of the Letter "i": Toward Writing Style Anonymization*. PETS 2012.

Morton A.Q. 1965. *The Authorship of Greek Prose*. Journal of the Royal Statistical Society (A), 128, 169-233.

A. Narayanan, H. Paskov, N. Gong, J. Bethencourt, E. Stefanov, R. Shin, and D. Song. 2012. *On the feasibility of internet-scale author identification.* Proceedings of the 33rd conference on IEEE Symposium on Security and Privacy.

Oakes M. 2005.*Statistics for Corpus Linguistics*. Edinburgh University Press.

F. Peng, D. Scuurmans, V. Keselj, and S. Wang. 2003. *Language Independent Authorship Attributions using Character Level Language Models*. Proc. of the 10th Conference of the European Chapter of the Associations for Computational Linguistics (EACL'03).

Marius Popescu, Liviu P. Dinu. 2007. *Kernel methods and string kernels for authorship identification: The Federalist Papers case.* Proceedings International Conference RANLP - 2007, pp 484-487.

Alan Ritter, Sam Clark, Mausam and Oren Etzioni. 2011. *Named Entity Recognition in Tweets: An Experimental Study*. Empirical Methods in Natural Language Processing, 2011.

Stamatatos E., Fakotakis N. and Kokkinakis G. 2001. *Computer-based authorship attribution without lexical measures*. Computers and the Humanities 35, pp. 193-214.

Stamatatos E. 2009. *A survey of modern authorship attribution methods*. J. Am. Soc. Inf. Sci. Technol., vol. 60, no. 3, pp. 538–556.

W. Teahan, R. McNab, Y. Wen, and I. H. Witten, 2000. *A compression-based algorithm for Chinese word segmentation*. Comput. Linguist., vol. 26, no. 3, pp. 375–393.

Tweedie F. J. and Baayen R. H. 1998. *How Variable May a Constant Be? Measures of Lexical Richness in Perspective*. Computers and the Humanities, 32 (1998), 323-352.

M. van der Velden, K. El Emam. 2012. *Not all my friends nee to know: A qualitative study of teenage patients, privacy and social media.* Journal of the American Medical Informatics Association, Published Online First, doi:10.1136/amiajnl-2012-000949.

Zhao Y. and Zobel J. 2007. *Searching with style: authorship attribution in classic literature*. Proceedings of 30th Australasian Conference on Computer Science, Vol. 62, pp. 59-68.

Zheng R., Qin Y., Huang Z., & Chen H. 2006. *A Framework for Authorship Analysis of Online Messages: Writing-style Features and Techniques.* Journal of the American Society for Information Science and Technology 57(3): (2006), 378-393.