Author attribution with email messages

Dominique Estival

University of Sydney, Australia
Dominique.Estival@usyd.edu.au

Abstract. This paper gives an overview of a research project which resulted in the development of a prototype tool for author attribution (TAT). The tool was trained on two email corpora (English and Arabic) and produces probabilities for the basic demographic traits (gender, age, geographic origin, level of education and native language) and some psychometric traits of the author of a text. The system also provides the probability of a match between a given email and other texts. I will describe the overall system and its components and outline the ways in which the email data is processed and analysed, before describing the Machine Learning setup used to produce the classifiers for the different author traits. I will conclude by presenting the experimental results, which are promising for most traits examined.

Keywords: author attribution, email analysis, machine learning.

1. Author attribution and author profiling*

Authorship attribution is the task of deciding for a given text which author (usually from a predefined set of authors) has written it. Classic examples include authorship attribution studies on the Bible [1], Shakespeare's works [2] or the Federalist Papers [3]. For a long time, the main applications were restricted to literary texts. Recently, authorship attribution has gained new life in the fight against cyber crime and in a more general search for reliable identification techniques [4, 5, 6, 7]. Being able to predict automatically the identity of authors from their texts has a number of potential applications. For example, if a text poses any type of threat, then identifying the source of the threat is the first step in countering it. In this context, author profiling forensics can be helpful to at least narrow the list of potential authors [8, 9, 10]. Another area where author identification and profiling can provide valuable information is in deriving marketing intelligence from the acquired profiles [11] and in the rapidly growing field of sentiment analysis and classification [12].

The task of authorship attribution has traditionally been carried out on data from small sets of authors. For larger data sets, involving more authors, the challenge of identifying individual authors is more difficult. In such cases, predicting characteristics, or traits, of authors can be a good alternative and provide clues as to the author's identity.

* This is a shorter version of a paper which has been submitted for publication. I wish to thank my colleagues at Appen who contributed to the success of the TAT project and who are co-authors of the full version of the paper: Tanja Gaustad, Ben Hutchinson, Son Bao Pham and Will Radford.

Author profiling is the task of predicting one or more such author traits and an author profile consists of the resulting set of one or more predicted traits. Importantly, and contrary to author attribution, the author profiling task is possible even when documents by the author are not in the training data. Also in contrast to author attribution, greater accuracy can be expected when the training data contains texts from more authors, because the models learned for each trait are then expected to be more robust.

In this talk, I will discuss some aspects of a project in which we developed a language-independent prototype system for text attribution. The Appen Text Attribution Tool (TAT) aims to provide information about the authors of texts for a variety of document types and a range of languages. The current implementation of the Appen TAT produces profiles for the authors of email messages written in English [13] and in Arabic [14]. These profiles consist of probabilities for the author's basic demographic traits such as gender, age, geographic origin, level of education and native language, as well as for some psychometric traits.

Most research into author profiling focuses on the prediction of a small number of traits, e.g. gender [9], gender and age [15], neuroticism and extraversion [9], neuroticism, agreeableness, extraversion and conscientiousness [12], neuroticism, agreeableness, extraversion, conscientiousness, and openness [16]. Our project, as far as we know, covers the largest number of traits to be predicted, since we predict a total of ten traits for English, five demographic and five psychometric traits, and seven traits for Arabic, three demographic and four psychometric traits.

Various machine learning techniques have been employed for profiling or trait prediction. Our approach has been to experiment with a number of machine learners and to select the best combination of machine learning algorithm and feature set for each trait.

For author profiling applications, an interesting question is how much data is actually necessary to perform reliable profiling. While we do not claim to be able to give a definite answer to this question, our experiments show that we can already get useful results with the relatively small amount of data we used for training.

2. The Data

As in all author attribution and author profiling studies, the choice of data was an extremely important issue. We decided to focus on email messages, as opposed to blogs or chat room data, and to collect spontaneous rather than artificially elicited data. The corpus we used was collected specifically for the Appen TAT project and the data collection itself was a large part of the overall effort for the project. Our corpus thus constitutes a completely new data set, consisting of two sets of emails from 1,033 English speakers and from 1,030 Arabic speakers.

We collected emails in several varieties of English, from both native and non-native speakers of English, coming from different geographical areas: on the one hand, native speakers of US English and native speakers of Australian or New Zealand English; on the other hand, native speakers of Spanish living in the US and native speakers of Egyptian Arabic living in Egypt. The Arabic data set consists of emails written by native speakers of Egyptian Arabic.

Table 1 gives an overview of the email corpus, with statistics for the number of

authors, number of emails and total number of words for each language. For the Arabic data, we also include the number of emails in Arabic script and in Latin script.

Collection	Native lang.	# authors	# emails	# words total	# words by author
USA	English	415	4,533	2,405,792	1,886,389
UK	English	23	273	178,400	137,238
AUS/NZ	English	133	1,387	513,065	437,454
USA	Spanish	174	1,823	519,504	461,767
Egypt	Arabic	288	1,820	451,903	444,325
Total English		1,033	9,836	4,068,664	3,367,173
Egypt	Arabic	1,030	8,028	2,153,333	2,153,333
Arabic script			7,267		
Latin script			761		
Total overall		2,063	17,864	6,221,997	5,520,506

Table 1. Overview of the email corpus

2.1. Corpus collection

The data collection processes for the English and Arabic data differed slightly in that the Arabic authors were asked to come to a central location where they were supervised, while the process used for the collection of the English data was completely on-line and unsupervised. However, in both cases, the process included notification of privacy and the assurance that the identity of the respondents would be protected. In both types of collection, the respondents agreed to fill out a web questionnaire in order to provide demographic and psychometric information about themselves and then to donate at least ten email messages. The demographic traits cover basic demographic information about the author: *age*, *gender*, *native language*, *level of education* and *main country of residence*. For the Arabic data collection, native language is always Arabic and country of residence is always Egypt. For the psychometric traits, all the English collections (with the exception of Egyptian English) use a short version of the International Personality Item Pool (IPIP) questionnaire [17]. The psychometric traits for the Arabic collection are based on a customized version of the short Eysenck Personality Questionnaire Revised (EPQR-S) [18]. While the IPIP yields five psychometric traits: *agreeableness*, *extraversion*, *neuroticism*, *conscientiousness*, and *openness* (also referred to as the "Big Five") [19], the EPQ [20] aims to analyse personality along four traits, namely *extraversion*, *neuroticism*, *psychoticism* (split up into *conscientiousness* and agreeableness in the Big Five) and *lie*.

After completing the questionnaire, respondents in the English data collection forwarded some of their previously sent email messages, e.g. from their email client's sent mail folder, to the data collection email address. In the Arabic data collection, the writers either forwarded previously sent email messages or composed new emails which they then sent to their recipients and forwarded to the data collection email address. In either case, the raw email messages were then stored on a dedicated mail server. The email messages were then normalised and validated.

2.2. Data validation

The email messages were first checked manually to filter out erroneous content such as foreign language emails or forwarded chain letters and to ensure consistency and accuracy of the documents in the corpus. As with any collection of email data, plagiarism and copying were issues that required careful checking of all the data received and we developed a plagiarism detecter to reject emails which had already been submitted. In addition to the minimum requirement of five lines per email message, data from authors who did not meet a set of satisfaction criteria removed. These criteria were: 1) a valid questionnaire received for a given author; 2) at least five valid email

messages for the author; and 3) a total word count for that author's valid email messages of at least 1000 words for English emails. Research has shown that the more complex morphology of Arabic (combined with a rich vocabulary) leads to a higher degree of inherent sparseness in Arabic data compared to similar English data. This suggests that larger amounts of data are needed for statistical Natural Language Processing (NLP) applications in Arabic [21]. Therefore, while the minimum amount of data we aimed to collect was set at 1000 words per writer for English, we decided to aim for 2000 words per writer for Arabic. This is of course a separate question from how much data is needed to perform author profiling in either English or Arabic. An additional requirement was that the emails be from different domains, such as personal or business emails. In the end, 2063 respondents were validated with a combined total of 17,864 email messages. The final corpus consists of only about 50% of the total number of emails collected.

3. The Appen TAT

The main goal of the project was to develop a tool which could provide analysts with information about the authors of documents. The requirements were that the tool should take documents as input, produce statistical descriptions of various characteristics of those documents and predict author traits as output. The intended users of the tools are analysts and investigators who are not experts in linguistics and who use documents as evidence in their investigations. The prototype delivered to the client allows users to 1) submit a document and retrieve the predicted author profile for that document; 2) retrieve documents with a similar author profile; 3) specify an author profile and retrieve documents matching that profile.

The tool can thus provide analysts with additional investigative information based on the texts they submit to the tool. Such information can help them identify individuals of interest and potentially link together separate investigations.

3.1. System description

Figure 1 gives a high-level overview of the Appen TAT system. It consists of several data repositories and a number of components for deriving features and for building classifiers.

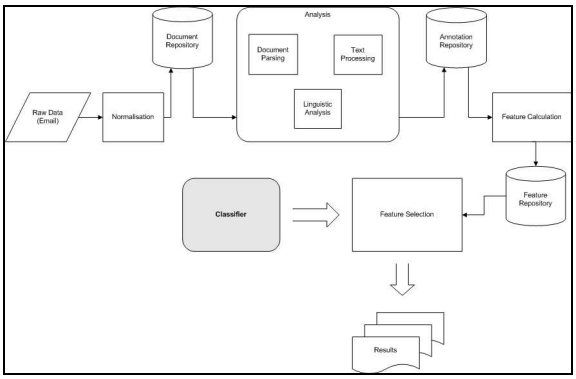

Fig 1. Appen TAT System Diagram

While the current Appen TAT prototype takes Arabic and English email input, the underlying processing architecture is language independent and can be extended to other types of documents and to other languages. The modular processing architecture is organized around a chain of modules, allowing flexible experimentation with different combinations of modules and providing a robust software framework which promotes reuse of modules and components. The analysis of a document is represented in stand-off annotations and saved in a common structure, the Annotation Repository.

During development of the Appen TAT, the data was first analysed and converted into features used by the classifiers for each of the author traits. The classifiers use both document and linguistic features (see Section 4). The classifiers in the operational system were chosen as the results of experiments with various Machine Learning algorithms and feature combinations (see Section 5).

3.2. The user interface

Developing the user interface for the Appen TAT was a significant part of the project. The interface was designed in collaboration with a professional designer and was first presented to the client for comments and suggestions. A usability study was then conducted with potential users of the tool. Their detailed comments and suggestions were taken into account for the final user interface delivered with the prototype, shown in Fig.2.

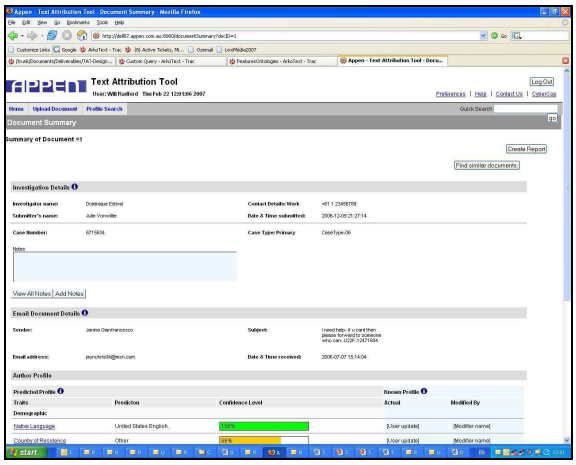

Fig 2. Appen TAT User Interface

4. Analysis

The analysis phase produces the annotations which are then used in the feature extraction phase. There are two ways annotations can be created: automatically, by the analysis modules described below, and manually by human annotators according to pre-specified guidelines using the Callisto tool [22]. Since the Appen TAT needs to be used by non-linguists, our goal is to produce automatically created features. However, during development, we relied on manual annotations for validation purposes and some of the modules described here were trained on manually annotated data. The analysis stage consists of three sets of modules: document parsing, text processing and linguistic analysis, each producing different types of annotations.

4.1. Document Parsing

Each line in the body text of an email document into 5 categories:

1. Author text: text that was written by the author and that is not contained in an embedded reply chain of email messages;

2. Signature: email signature text, which typically includes contact information, professional details, and/or quotations;

3. Advertisement: text automatically appended by the author's email client, such as Yahoo and Hotmail;

4. Quoted text: extended quotations, e.g. song lyrics, poems, newspaper articles;

5. Reply lines: text that was written in a previous email message that the author is either forwarding or replying to, including text by other writers, text in previous emails by the author of the current email, with their email signatures, advertisements and quotations.

Document parsing is a crucial stage, as it is the linguistic features of the author text which provide most of the clues for author attribution. The input to document parsing is an email

document and the output is the same email document, with the lines of the body text categorised into one of those five categories.

We experimented with Jangada [23], a tool which identifies signature blocks and replies, but unfortunately it did not perform very well on our data, so we developed our own document parser. To compare the performance of Jangada with our document parser, we used ten-fold cross validation on all of the English data. For each train-test partition, our document parser was trained on the training part and its performance was compared with Jangada's on the testing part. As Jangada can only identify Author text, Reply and Signature lines, we cannot compare the performances of the two tools in recognising all 5 categories. However for the task of identifying those 3 categories, our document parser achieved an F-score of 88.16%, while Jangada performed at 64.22%. For the task of identifying only author lines, our document parser reached an F-score of 90.76%, compared to 74.64% from Jangada.

4.2. Text and Linguistic Processing

In the general case, i.e. for both English and Arabic, text processing consists of two stages: segmentation and punctuation analysis. First, the text in the email is split into paragraphs and paragraphs are then split into sentences and tokens. The latter task is performed with third party tools [24] which generate *paragraph*, *sentence* and *token* annotations respectively. The use of sentence punctuation marks and other special characters is then analysed. The special marks and characters include, but are not limited to, special markers, e.g. two hyphens \- -" followed by a newline which often indicate that an email signature follows; quotation marks, which sometimes signal the presence of a quotation; and emoticons, such as \:-)" or \:o)". This information is stored as attributes of *token* annotations which are used for calculating character-level features.

The aim of the linguistic analysis stage is to produce more linguistically informed annotations, such as Part-Of-Speech (POS) tags. Unlike text processing, linguistic processing deals with aspects of texts which are usually language-dependent and thus requires linguistic resources such as word lists.

To identify certain key phrases, we developed a Named Entity Recognizer (NER) using gazetteers and grammars. We decided to implement our own NER to identify people, locations, organisations, dates etc. because most available systems were developed on news corpora, i.e. a very different domain from ours. All the heuristics in our NER are based on email data; additional lexicons were developed manually to identify set phrases, for instance farewells and greetings.

Arabic emails present a number of challenges for NLP and specific modules had to be developed to handle these, in particular: different ways of writing Arabic in Latin script (so-called “franco-arabic”), spelling variants in the Egyptian dialect and possible spelling normalisation, morphological complexity, the use of English loanwords and their transliterations, spelling errors and typos.

5. Machine Learning Classification

Many problems in NLP have lent themselves to solutions using statistical language processing techniques. Author profiling can be considered a type of document classification task, where the classes correspond to traits of the authors. These traits are arranged along various dimensions, with different options for each dimension being mutually exclusive. For example *male* and *female* are the

possibilities for the *gender* dimension. For each dimension, the email and questionnaire data are used to construct classifiers, using a range of ML techniques.

A document constitutes a single data instance. For each experiment ten-fold cross-validation was used and we also used ten-fold cross-validation during training, for feature selection and model parameter tuning. Once the best combination of ML classifiers, parameters and feature selection was determined, that model was used to classify the test data to evaluate the performance of the chosen model.

For each author, a feature vector is calculated. Typically, a feature is a descriptive statistic calculated from both the raw text and the annotations. For example, a feature might express the relative frequency of two different annotation types (e.g. number of words/number of sentences), or the presence or absence of an annotation type (e.g. signature). For the English data, 689 features were calculated. These were divided into three main groups, namely character-level, lexical, and structural features. For the Arabic data, 518 features were calculated, also divided into several subgroups.

The aim of the classifier is to match feature vectors from the document with author traits. Ordered pairs of feature vectors and author traits are used to train and tune machine learning classifiers. Formally, classifiers are functions which map feature vectors to author traits and there will be classifiers for each author trait such as *gender*, *age*, etc. We experimented with various machine learning algorithms as classifiers, using the WEKA toolkit [25] to find the best classifier for each trait. During training, classifiers are created by the selection of sets of features for each author trait, and classifier parameters are tuned through cross-validation. To evaluate and test the classifiers, new documents are given as input and existing classifiers are selected to predict author traits.

The machine learning algorithms we tried include decision trees (J48 [26], RandomForest [27]), lazy learners (IBk [28]), rule-based learners (JRip [29]), Support Vector Machines (SMO [30]), LibSVM [31]), as well as ensemble/meta-learners (Bagging [32]), AdaBoostM1 [33]). These algorithms were used in combination with feature selection methods based on either a feature sub-set evaluator together with a search method (consistency subset evaluator with a best-first search) or a single attribute evaluator with various numbers of attributes selected (X^2, GainRatio, and InformationGain) (see chapter 10.8 in [25] for details).

6. Results and Discussion

Table 2 shows the best results for English on all ten traits (demographic and psychometric) whereas Table 3 shows the best results for Arabic on the seven traits to be predicted. Both tables also include the baselines associated with each separate classification task, calculated on the corresponding data sets.

Trait	ML algo.	Feat. sel.	Best Features	Results	Baseline
Age	SMO	–	all	56.46	39.43
Gender	SMO	–	all	69.26	54.48
Language	RandForest	InfoGain	all-correlate	84.22	62.90
Education	Bagging	–	all-funcWord	79.92	58.78
Country	SMO	–	all	81.13	57.29
Agreeableness	IBk	–	char+struct	53.16	40.51
Conscient.	IBk	–	char+struct	54.35	43.72
Extraversion	LibSVM	–	char+struct	56.73	45.17
Neuroticism	IBk	–	char+struct	54.29	42.34
Openness	RandForest	–	struct	55.32	47.28

Table 2. Results for the English data

Trait	ML algo.	Feat. sel.	Best Features	Results	Baseline
Age	Bagging	InfoGain	all-arabicLex	72.10	70.09
Gender	SMO	–	all	81.15	72.16
Education	Bagging	InfoGain	all	93.66	93.62
Extraversion	SMO	–	all-arabicMorph	54.35	48.27
Lie	Bagging	InfoGain	all	52.30	40.41
Neuroticism	Bagging	InfoGain	all	54.93	43.42
Psychoticism	Bagging	InfoGain	all	56.98	49.39

Table 3. Results for the Arabic data

These results show that, for the demographic and the psychometric author profiles, classification is significantly improved over the baseline for all ten traits in the case of English and for six out of seven traits for Arabic. This demonstrates that the approach we took of combining ML algorithms, together with our particular feature set, is successful for binary as well as n-ary classifications on very diverse classification tasks. For predicting level of education in Arabic emails, virtually no improvement can be seen over the baseline; this is due to the extremely skewed data, as indicated by the very high baseline of 93.62%. Even though the baselines for the other Arabic demographic traits are also quite high, our system still achieves a better classification accuracy for age and gender than the majority baseline.

The results of experiments aimed at discovering how well a range of ML algorithms perform on this data set for various demographic and psychometric author traits show that the chosen approach works well for author profiling and that using different classifiers in combination with a subset of available features can be beneficial for predicting single traits.

7. Conclusion

I have presented a research project in which we implemented machine learning techniques to classify a new set of data in order to produce author profiles based on a number of author traits. Such a project combines software engineering, data collection and corpus studies, computational linguistics, machine learning experiments and interface design. It is only possible with a team of people with diverse qualifications and skills and, as with the design of the user interface, may require contracting special professional skills.

Such a project also requires flexibility in the choice of tools and techniques. We experimented with a number of third party tools before deciding whether to integrate them (e.g. open source ML tools), re-implement them (e.g. Named Entity Recogniser) or even design a new tool (e.g. Document Parser). For document processing and linguistic analysis, we used well-known tools to handle English but had to create new modules to handle Arabic and the specific problems of Arabic email texts.

Future research will need to investigate deeper features, such as syntactic information or writing style, which might help to classify the author traits more accurately. It would also be interesting to identify more specialised feature sets for each author trait.

Acknowledgements

This research was carried out within the framework of a US Government BAA grant (IS-QD-2467) with joint funding from Appen and the US Government. I want to acknowledge and thank all our colleagues at Appen who contributed to the success of the project and especially Judith Bishop who managed the data collection.

References

[1] R. Friedmann. (1997) *Who wrote the Bible?* San Francisco: Harper.

[2] G. Ledger and T. Merriam. (1994) “Shakespeare, Fletcher, and The Two Noble Kinsmen”. Literary and Linguistic Computing, 9(3): 235-248.

[3] F. Mosteller and D.L. Wallace. (1964) “Inference and Disputed Authorship: The Federalist.” Series in behavioral science: Quantitative methods edition. Addison-Wesley.

[4] A. Abbasi, H. Chen. (2005a) “Applying authorship analysis to Arabic web content”. In P. B. Kantor et al. (eds.), Intelligence and Security Informatics, Proceedings of the IEEE International Conference on Intelligence and Security Informatics (ISI 2005), pp. 183-197. Berlin: Springer.

[5] O. de Vel, A. Anderson, M. Corney, G. Mohay. (2001) “Mining email content for author identification forensics”. SIGMOD Record, 30(4): 55-64.

[6] O. de Vel, A. Anderson, M. Corney, G. Mohay. (2002) “E-mail authorship attribution for computer forensics”. In D. Barbara and S. Jajodia (eds.), Data Mining for Security Applications. Kluwer Academic Publishers.

[7] R. Zheng, Y. Qin, Z. Huang, H. Chen. (2003) “Authorship analysis in cybercrime investigation. In H. Chen et al. (ed.), Proceedings of the First NSF/NIJ Intelligence and Security Informatics Symposium, pp. 59-73. Springer.

[8] M. Corney, O. de Vel, A. Anderson, G. Mohay. (2002) “Gender-preferential text mining of e-mail discourse”. In Proceedings of the 18th Annual Computer Security Applications Conference (ACSAC 2002), pp. 282-292. Las Vegas.

[9] S. Argamon, S. Dhawle, M. Koppel, J. W. Pennebaker. (2005) “Lexical predictors of personality type”. In Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America. St. Louis.

[10] A. Abbasi, H. Chen. (2005b) “Applying authorship analysis to extremist-group web forum messages”. IEEE Intelligent Systems, 20(5): 67-75.

[11] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, T. Tomokiyo. (2005) Deriving marketing intelligence from online discussion”. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pp. 419-428. Chicago.

[12] J. Oberlander, S. Nowson. (2006) “Whose thumb is it anyway? Classifying author personality from weblog text”. In Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pp. 627-634. Sydney.

[13] D. Estival, T. Gaustad, S. B. Pham, W. Radford, B. Hutchinson. “Author profiling for English emails”. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*. Melbourne, Australia. pp. 263-272.

[14] D. Estival, T. Gaustad, S. B. Pham, W. Radford, B. Hutchinson. “TAT: an author profiling tool with application to Arabic emails”. In *Proceedings of the 5th Australasian Language Technology Workshop*, Melbourne, Australia. pp. 21-30.

[15] M. Koppel, J. Schler, S. Argamon, E. Messeri. (2006) “Authorship attribution with thousands of candidate authors”. In Proceedings of SIGIR, pp. 659-660.

[16] F. Mairesse, M. Walker. (2006) “Words mark the nerds: Computational models of personality recognition through language”. In Proceedings of the 28th Annual Conference of the Cognitive Science Society (CogSci 2006), pp. 543-548. Vancouver.

[17] T. Buchanan, J. A. Johnson, L. R. Goldberg. (2005) “Implementing a Five-Factor personality inventory for use on the internet”. European Journal of Psychological Assessment, 21:115-127.

[18] L. Francis, C. Lewis, H.-G. Ziebertz. (2006) “The short-form revised Eysenck personality questionnaire (EPQR-S): A German edition”. Social Behaviour and Personality, 34(2): 197-204.

[19] W. T. Norman. (1963) “Toward an adequate taxonomy of personality attributes: replicated factor structure in peer nomination personality rating”. Journal of Abnormal and Social Psychology, 66:574-583.

[20] H. Eysenck, S. Eysenck. (1975) *Manual of the Eysenck Personality Questionnaire.* London: Hooder and Stoughton Educational.

[21] A. Goweder, A. De Roeck. (2001) “Assessment of a significant Arabic corpus”. In Proceedings of the ACL/EACL Workshop on Arabic Language Processing: Status and Prospects. Toulouse.

[22] Mitre. (2006) Callisto. http://callisto.mitre.org/. Version 1.4.0.

[23] V. Carvalho, W. Cohen. (2004) “Learning to extract signature and reply lines from email”. In Proceedings of the Conference on Email and Anti-Spam (CEAS-2004). Mountain View.

[24] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. (2002) “GATE: A framework and graphical development environment for robust NLP tools and applications”. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 168-175.

[25] I. H. Witten, E. Frank. (2005) *Data Mining: Practical machine learning tools and techniques.* San Francisco: Morgan Kaufmann, 2nd edition.

[26] R. Quinlan. (1993) *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann.

[27] L. Breiman. (2001)” Random forests”. Machine Learning, 45(1): 5-32.

[28] D. Aha, D. Kibler, M. Albert. (1991) “Instance-based learning algorithms”. Machine Learning, 6(1): 37-66.

[29] W. Cohen. (1995) “Fast effective rule induction”. In Twelfth International Conference on Machine Learning, pp. 115-123.

[30] S. Keerthi, S. Shevade, C. Bhattacharyya, and K. Murthy. (2001) “Improvements to Platt's SMO algorithm for SVM classifier design. Neural Computation, 13(3): 637-649.

[31] C.-C. Chang, C.-J. Lin. (2001) *LIBSVM: a library for support vector machines.* Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[32] L. Breiman. (1996) “Bagging predictors”. Machine Learning, 24(2):123-140.

[33] Y. Freund, R. Schapire. (1996) “Experiments with a new boosting algorithm”. In Thirteenth International Conference on Machine Learning, pp. 148-156. San Francisco.