Selecting Systemic Features for Text Classification

Casey Whitelaw and **Jon Patrick**
Language Technology Research Group
Capital Markets Co-operative Research Centre
School of Information Technologies
University of Sydney
{*casey, jonpat*}*@it.usyd.edu.au*

Abstract

Systemic features use linguistically-derived language models as a basis for text classification. The graph structure of these models allows for feature representations not available with traditional bag-of-words approaches. This paper explores the set of possible representations, and proposes feature selection methods that aim to produce the most compact and effective set of attributes for a given classification problem. We show that small sets of systemic features can outperform larger sets of word-based features in the task of identifying financial scam documents.

1 Introduction

Text classification is among the most widespread applications of computational linguistics. The range of services that offer text classification continue to grow, and include such mainstream applications as email and web content filtering. The classification of documents by machine learning techniques requires a representation of each document as a set of features; almost without exception, these features are based on the presence, absence, or frequency of words in the text. This 'bag-of-words' model is popular both due to its ease of implementation, and its excellent performance on many tasks. Topic-based classification, such as newswire or newsgroup tasks, is well-suited to this automated keyword-spotting approach. These are cases in which the presence of a topic-related word such as 'wheat' is a very strong indicator of a document's class.

The bag-of-words model makes large simplifying assumptions about a document. It assumes that there is no textual structure; no ordering of paragraphs in the text, sentences in a paragraph, clauses in a sentence, or words in a clause. In addition, it assumes that the occurrence of each word is independent of each other word. These assumptions, in providing a much simpler picture of the document, destroy much of the text's meaning. Work has been done to restore this information using semantic resources such as WordNet (Scott and Matwin, 1998) or using syntactic information (Carr and Estival, 2002). There is also growing interest in classifying texts on non-denotational meaning, such as writing style, authorship identification (van Halteren, 2004) and sentiment analysis (Pang and Lee, 2004). These new areas highlight the properties of a document that are currently slipping through the cracks.

This paper takes another approach to providing a better representation than bag-of-words. In line with Systemic Functional Linguistic theory, the words of a text are treated as evidence of semantic choices being made by the author. These choices form systems, and each document is modelled as the set of choices it makes within these systems. This knowledge of the semantic relationships between features allows for more sophisticated representations that more accurately capture characteristic linguistic differences. In Section 2 we enumerate these representations and discuss how their semantics differ. Section 3 describes the Scamseek project and the use of systemic features in the identification of financial scams. The results provided in Section 4 show that smaller number of systemic features can outperform larger numbers of word-based features in text classification.

2 Systemic Features

Systemic Functional Linguistics (SFL) is a linguistic theory that approaches language as a social resource for meaning-making (Halliday, 1994). Language is not seen as a collection of discrete phrase production rules working upon a deeper syntactic structure, but as an interwoven collection of systems realising a deeper semantic structure and functional intention. SFL explicitly deals with three types of meaning (metafunctions) in text: the *ideational* (content of the text), the *textual* (organisation of the text), and the *interpersonal* (social positioning of the text) meanings, each of which contribute to the formation of a document.

SFL uses *system networks* as a way to represent the patterns of language choice related to a particular meaning. A system network is defined both graphically and algebraically (Matthiessen, 1995) as a hierarchy of choices: at the most delicate level, these choices result in particular lexical or grammatical artifacts. SF linguists have proposed standard system networks for most aspects of the English language.

SFL has been applied in natural language processing since the 1960s, but has been adopted most widely within the field of text generation (Matthiessen and Bateman, 1991). Most recently, systemic analysis has been used with machine learners in more statistical NLP tasks such as functional clause classification (O'Donnell, 2002). The increased interest in attitude and affect has also seen SFL's theory of appraisal used to augment sentiment classification (Taboada and Grieve, 2004).

Systemic features are a way to describe the usage of a system network within the document as a whole. Systemic features were introduced as a way of identifying the interpersonal distance of documents (Whitelaw et al., 2004), using only a single system network. Features from multiple system networks have been used together to classify different styles of academic writing (Argamon and Dodick, 2004). These types of grammar models have been shown to be well suited to the task of describing the non-denotational or stylistic properties of writing (Whitelaw and Argamon, 2004).

2.1 Types of System Networks

Two types of system network are used in this paper, both constructed using Systemic Functional Linguistic theory. The first, grammar models, are based on the general linguistic descriptions provided in linguistics texts, eg. (Matthiessen, 1995), and are similar to those used previously for stylistic text classification. The specific systems used here include:

- CONJUNCTION: models how clauses expand on their context through elaboration (*that is*), extension (*moreover*), or enhancement (*then*, *next*).
- PRONOMINAL/DETERMINATION: models the way in which referents are identified in a text. This system has been used to classify texts on the basis of interpersonal distance (Whitelaw et al., 2004)
- COMMENT: describes the status of a clause within the context as eg. evaluative/judging (*sensibly*), desiderative (*unfortunately*), or assertive (*certainly*).
- MODALITY: is a rich system that describes the likelihood (*probably*), frequency (*might*), and necessity (*should*) of events.

Grammatical models such as these provide a general profile of language use within a document and a register. An advantage of these general models is their domain independence; the distinctions made within these systems are based on the manner in which the document was written, rather than its topic. Manual linguistic research has given evidence that scam documents differ from normal documents in their language (Herke-Couchman, 2003), and so it is expected that features from these systems will assist in this classification.

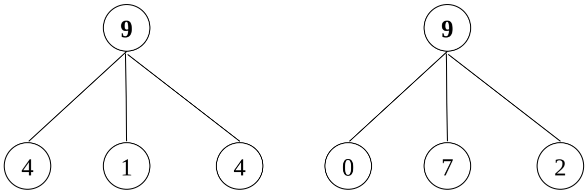


Figure 1: Aggregating counts smooths differences at greater delicacy

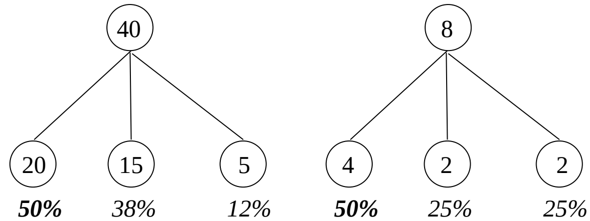


Figure 2: Proportional features are a local and size-independent measure

Register models, in contrast to the general applicability of the grammatical models, describe specific linguistic traits that are characteristics of individual registers. Register models are compiled manually by trained SF linguists based on their analysis of a training corpus. A register, in SFL terminology, is a group of texts whose language selections vary from the general language system in similar ways; a skewing 'of probabilities relative to the general systemic probabilities' (Matthiessen, 1993). In the absence of a fully developed system network for English, register models each define portions of language use that are characteristic and discriminatory within the current classification task.

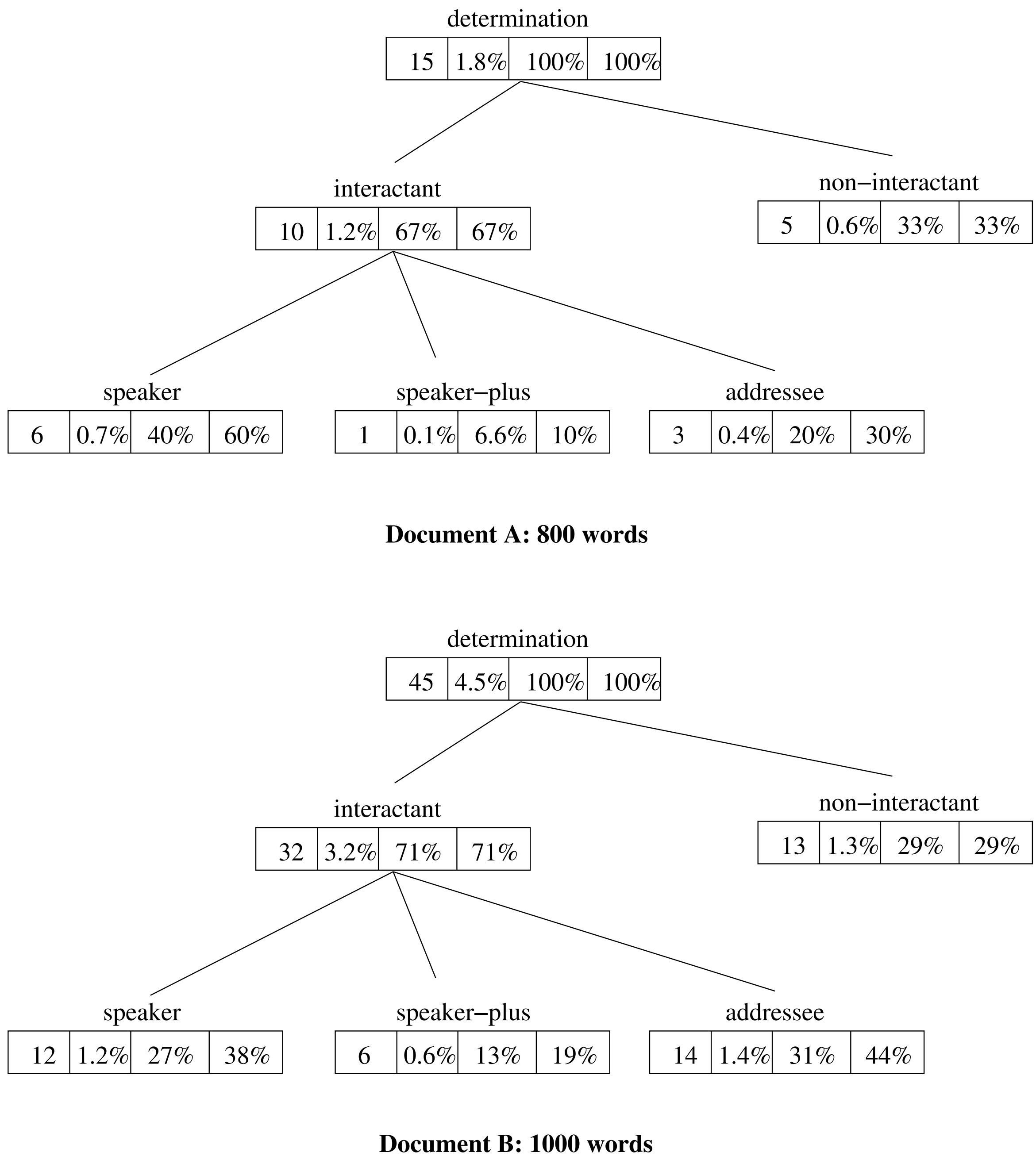
2.2 Leveraging Systemic Structure

In a standard 'bag-of-words' approach, the contribution of a word to a document is given by its relative frequency; how rarely or often that word is used. This implicitly uses a language model in which all words are independent of each other. Crucially, this does not and cannot take into account the choice between words, since there is no representation of this choice. Placing words within a system network provides a basis for richer and more informative feature representation. There are two main advantages to be gained from systemic information.

Firstly, it allows for categorical features that are based on semantically-related groups of words, at all levels in the network. By collecting aggregate counts, individual variations within a category are ignored. Figure 1 shows the raw counts of the same system in two documents; at the lower level, closer to lexis, the distributions of counts are highly dissimilar. At the higher level, these differences have been smoothed, and the documents look the same. This aggregation also helps alleviate the problems associated with representations containing large numbers of very sparse features.

For a given register, it may be the case that important and characteristic language choice occurs at a very fine level, distinguishing between usage of individual words. This word-level information is kept intact, as in a bag-of-words approach. In another register, it may be the usage of a category, such as interactant, that is characteristic. The usage of any words within the category may appear random while maintaining consistent category usage. These higher-level features are not available in a traditional bag-of-words approach, hence these patterns may be lost as noise.

The second and more important difference to traditional feature representation is the representation of language choice. SF theory treats language use as a series of selections within systems; at any point in the system network, or tree as it has been modelled here, the selection is restrained to the immediate sub-systems. The choice is not between one word and any other, or even one system and any other, but a series of semantically-driven choices within the system. A bag-of-words model can model only choice between one word and any other; a choice between arbitrary words such as 'dog' and 'elegant'. Comparative features such as these can only be used within an appropriate theory-driven structure, which is provided here through the use of SFL and system networks. Figure 2 shows the potential for comparative features to reveal similarities not immediately apparent in a text. The leftmost node in each system contributes 50% to parent system usage, despite markedly different numbers of occurrences.



count	*term frequency*	*system %*	*system contribution*

Figure 3: Different feature representations portray a text differently

2.3 Representing Systemic Features

Figure 3 shows a portion of the DETERMINATION system for two documents of different sizes, belonging to the same register. Four possible feature representations are given: from left to right, each node shows the total count, term frequency, system percentage, and system contribution. Each feature representation captures a different aspect of system usage in a document and register.

Raw counts (term frequency) (first column). The summed feature count, shown in the leftmost column, presents these two documents as highly dissimilar. Note also that this is only the top portion of the system, and that multiple levels exist below those shown. Raw term counts are usually not used directly as features, as they are heavily influenced by document length.

Document percentage (second column) is the standard basis for bag-of-words representations; it gives the proportion of the document accounted for by this term. It is commonly used since it normalises for document length; most topic-based document classification rightly assumes that the document length is not important (Sebastiani, 2002). In creating features for each sub-system, this representation can still take advantage of the aggregation and smoothing provided by the system, but does not take further advantage of the known structure.

System percentage (third column) gives the proportion of total system usage made up by this sub-system. In Document A, addressee occurs three times from a total of fifteen occurrences of determination in the document, giving it a system percentage of 20%. Within a document, system percentage is directly proportional to term frequency, but is independent to system *density* in the document. If another 800 words were added to Document A, but no more uses of DETERMINATION, the term frequency for a feature would halve while the system percentage remained constant. This makes it a suitable representation where distinctions are made not on how often a feature occurs, but the manner of its use. The system percentage of *speaker* is higher in Document A than Document B, despite higher term frequency in the latter. System percentage is also useful when the area of interest is a constant-size subsection of a variable-length document.

System contribution (fourth column) shows the ratio of sub-system to super-system occurrence. Again in Document A, speaker occurs six times and its super-system, interactant, occurs ten times, giving a system contribution of 60%. This is a strictly local measure of usage, and captures most directly the systemic notion of choice: once the decision to use a given super-system has been made, how often was this sub-system chosen as the realisation? This is a relative feature, and as such is independent of document length, total system usage, and usage of other portions of the system (see Figure 3). Despite the differences in lower-level choices, and in the raw counts of system usage, the system contribution of interactant in Documents A and B are very similar.

System contribution is not proportional or strongly correlated to document percentage, and the two measures provide useful and complementary information. Within a system instance, document percentage can be used to report the frequency not just of terms but of systems as well. System contribution does not capture how often a system is used, but rather its usage in relation to the other possible choices. In the same way as a register may be characterised by choice, it may also be characterised by frequent usage of a particular system, which will be highlighted by system percentage. The four complementary representations given here may each be useful in discerning characteristic system usage.

In implementing these representations, it is worth noting that not all system contribution features are necessary, and some can be removed. Features from a node which is an only child do not add information since there is no choice. In a system with a binary choice, either one of the features may be discarded since they have unit sum. Both system percentage and system contribution are meaningless at the root level, and system percentage and system contribution are identical at the first level below the root. These feature reductions can be performed deterministically before any further feature selection.

By mapping only the relevant portions of a document's meaning, systemic features also have the potential to increase computational efficiency by reducing the number of attributes used in machine learning systems, in comparison to broader bag-of-words methods. This should produce smaller feature sets with equal or better performance.

2.4 Selecting Systemic Features

We have presented four potential feature representations for systemic features. Depending on the behaviour of a system network in a particular classification task, the most appropriate representation may vary. In addition, the best feature type may change within a single system. We propose a simple feature selection method for systemic features.

For a given task, the attribute significance of each possible feature representation can be measured using a method such as information gain. By ranking the options for a single node upon an information metric, the best feature type for each node can be selected. This reduces the number of features, reduces the chance of performance loss through correlated features, and should combine the strengths of each feature type.

3 Scamseek: Identifying Financial Scams

We tested this range of possible systemic feature representations using models and data compiled as part of the Scamseek project[1]. Scamseek aims to identify a variety of criminal financial scams on the internet, using a combination of automatically and linguistically derived criteria.

The entire Scamseek corpus, collected and manually classified by ASIC experts, contains 7556 documents in a total of 58 registers. These registers fall into four broader classes which group financial scams, other scams, legitimate financial documents, and all other web pages. This coarser classification is of the most interest to the client, as potential scams are investigated regardless of scam type. For these experiments we used 1896 documents from 22 registers with a minimum of 20 documents per register.

As well as existing grammar models, a register model was developed by SF linguists for each of the Scamseek registers. These were treated in two ways.

[1]Scamseek is a joint project funded by The University of Sydney, the Capital Markets Cooperative Research Centre, the Australian Securities & Investments Commission and Macquarie University.

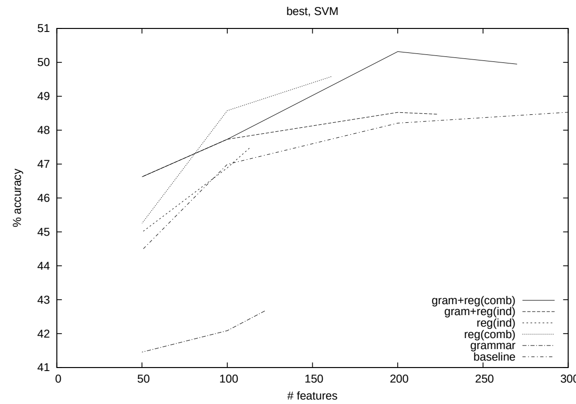


Figure 4: Results for each set of models, selecting the best feature at each node

Each register model can be considered as a selection from a full systemic description of English. Taken individually, a register model aggregates all topic- and genre-specific features, producing an overall picture of the 'topicality' of a text. The register models can also be combined to form a single system network, which is more complete but not topic-specific. In this case, it should function more like a grammar model in that the relative usage of systems should become more important. Both of these options were tested.

As well as testing grammar models and register models independently, the two types of system networks were combined. In all cases, features with no variation or no occurence in the corpus were removed. The systemic features were extracted from documents using an efficient partial parsing method (Whitelaw and Argamon, 2004). Each of the feature representation methods given in Section 2.3 were tested individually and in combination ('all'). The best-feature-per-node feature selection method ('best') was also tested for each feature set.

As a baseline, we used a bag-of-words representation using all of the words and phrases included in all the grammar and register models. Each feature set was tested at various sizes, using information gain to select features. Tests were performed using ten-fold cross-validation and the support vector machine (SVM) (Platt, 1998) im-

	gram	reg (ind)	reg (comb)
term frequency	24%	19%	14%
document %	20%	**36%**	**30%**
system %	18%	18%	27%
sys. contribution	**38%**	27%	29%

Table 1: Proportion of each feature type selected for 'best' sets

plementation used in the WEKA machine learning environment (Witten and Eibe, 1999) [2]

4 Results

Figure 4 shows the results from selecting the best combination of feature types. This also shows the best overall result achieved at 50.4%, using 200 features selected from the grammar and combined register models. This outperforms the full baseline result by two percent. The grammar models alone perform much lower than any of the register models, which is to be expected on this register-based classification.

Table 1 shows what types of features were selected in the best-feature-per-node process. As expected, the densely populated grammar models select more system contribution features. When register models are used individually they are very sparse, and there is less benefit from including relative features. In this case, document percentage makes up 36% of the feature set. The combined register model, which forms a single more fully-specified system network, selects equally from all feature types except term frequency. All representations were used by all models, and it is through this corpus- and system-specific selection that the best combination of feature types is found.

The relative performance of each feature type can be seen in Figure 5. As in most text classification, raw counts do not work as well as normalised features. Including all features from all nodes regardless of potential correlations, as shown by the solid line, produces worse results than using only the best combination of features.

As discussed in Section 2.2, features higher in a system network aggregate and smooth the features below it. When it is the use of semantic *categories*

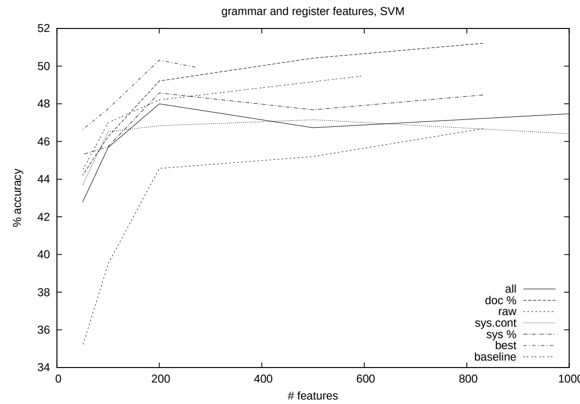


Figure 5: Results for each feature type

	features	accuracy
baseline	594	82.2%
grammar	200	82.4%
combined	200	84.4%
Scamseek	> 5000	$> 90\%$

Table 2: Class-based accuracy results.

of words that is important, these internal features will be favoured over lexis. This is the case for all the models tested: of the top hundred features in grammar models, 83 are internal. Experimental results bear out the advantage, with better performance for systemic features than the lexis-only baseline when both use document percentage.

Table 2 shows the class-based accuracy results for the best feature sets obtained. Registers are more similar to other registers in the same class, resulting in much higher performance than when classifying by register. The best set of 200 systemic features performed 2% better than the baseline bag-of-words system. Grammar models also outperformed the baseline despite poor register accuracy. This is evidence of the stylistic differences between these categories. The full Scamseek system, which combines bag-of-words features with more systemic features and other processing such as entity recognition, uses many more features and achieves much higher performance.

5 Conclusions

A document is more than a bag of words. As the forms of document analysis and classification con-

[2] Each experiment was also run using J48 decision trees and Naive Bayes, but produced consistently poorer results.

tinue to expand beyond topic detection, we must move towards a richer representation of a document. SFL provides one such linguistic model, and the representation of system models as features presented here shows the efficacy of a theoretically motivated approach. Systemic features allow for the production of smaller, denser feature sets that contain more sophisticated features than traditional methods. Grammar models can help build stylistic profiles of texts; register models supplement these with genre-specific linguistic phenomena. Through their combination, and a combination of new feature representations such as system contribution and system percentage, we have shown increased performance on the difficult task of identifying financial scams.

The system networks used in this research are still heavily tied to lexical realisations. The systemic feature extraction process can be efficiently expanded to include morphosyntactic and simple grammatical relationships; this will allow for the description of linguistic phenomena related to the logogenesis or unfolding of a text, such as the relative ordering of features. As more system networks are constructed for core sections of English SFL grammar, these models will be beneficial to a wide range of tasks including the classification of style, sentiment, attitude and affect.

References

Shlomo Argamon and Jeff T. Dodick. 2004. Linking rhetoric and methodology in formal scientific writing. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*.

Oliver Carr and Dominique Estival. 2002. Text classification of formatted text documents. In *Proceedings of the 2002 Australian Natural Language Processing Workshop*.

Michael A. K. Halliday. 1994. *Introduction to Functional Grammar*. Edward Arnold, second edition.

M. A. Herke-Couchman. 2003. Arresting the scams: Using systemic functional theory to solve a hi-tech social problem. In *ASFLA03*.

C. M. I. M. Matthiessen and J. A. Bateman. 1991. *Text generation and systemic-functional linguistics: experiences from English and Japanese*. Frances Pinter Publishers and St. Martin's Press, London and New York.

C. M. I. M. Matthiessen. 1993. Register analysis: theory and practice. In *Register in the round: diversity in a unified theory of register*, pages 221–292. Pinter, London.

Christian Matthiessen. 1995. *Lexico-grammatical cartography: English systems*. International Language Sciences Publishers.

M. O'Donnell. 2002. Automating the coding of semantic patterns: applying machine learning to corpus linguistics. In *Proceedings of the 29th International Systemic Functional Workshop*. University of Liverpool.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL 2004, Main Volume*, pages 271–278, Barcelona, Spain, July.

J. Platt, 1998. *Advances in Kernel Methods - Support Vector Learning*, chapter Fast Training of Support Vector Machines using Sequential Minimal Optimization. MIT Press.

Sam Scott and Stan Matwin. 1998. Text classification using WordNet hypernyms. In Sanda Harabagiu, editor, *Use of WordNet in Natural Language Processing Systems: Proceedings of the Conference*, pages 38–44. Association for Computational Linguistics, Somerset, New Jersey.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

M. Taboada and J. Grieve. 2004. Analyzing appraisal automatically. In *AAAI Spring Symposium of Exploring Attitude and Affect in Text*. AAAI.

Hans van Halteren. 2004. Linguistic profiling for authorship recognition and verification. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 199–206, Barcelona, Spain, July.

Casey Whitelaw and Shlomo Argamon. 2004. Systemic functional features in stylistic text classification. In *Proceedings of AAAI Fall Symposium on Style and Meaning in Language, Art, and Music*. AAAI.

Casey Whitelaw, Maria Herke-Couchman, and Jon Patrick. 2004. Identifying interpersonal distance using systemic features. In *Proceedings of AAAI Workshop on Exploring Attitude and Affect in Text: Theories and Applications*. AAAI Press.

Ian H. Witten and Frank Eibe. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.