

The Scamseek Project – Text mining for Financial Scams on the Internet

Jon Patrick

Sydney Language Technology Research Group
School of Information Technologies

University of Sydney
and

Capital Markets Co-operative Research Centre

jonpat@it.usyd.edu

Keywords: scam, Scamseek, systemic functional grammar, ontology

Abstract

The Scamseek project, as commissioned by ASIC, has the principal objective of building an industrially viable system that retrieves potential scam candidate documents from the Internet and classifies them as to their potential risk of containing an illegal investment proposals or advice. The project was operated in two stages over 15 months and produced multiple classifiers for different types of data and achieved higher than expected performance statistics on classifications, and was completed on time and under budget. The development of the system required the solution of two major problems in document classification, namely accurate identification of classes with very small footprints, <.1%, and classification using meaning intention rather than word strings. The approach taken uses the semantic model of language, Systemic Functional Grammar, to model the semantics of the scam classes and uses unigrams with significant language pre-processing to assist in separating irrelevant documents. This description is not complete in that some information is restricted for release due to security obligations. Prosecutions have been initiated by ASIC from classifications made by the very first production run of the system¹. ASIC operates the system on a 24/7 basis. The estimate of savings in human effort in its monitoring role is the order of 100 to 1. The estimate in savings to the community by bringing speedier detection and intervention of scams cannot be estimated readily but is likely to be of the order of tens of millions of dollars.

Scamseek Project Specifications

The Scamseek project was devised in two stages. The first stage had the aim of producing a production system for retrieving and classifying web pages. The second phase ran for 9 months to 30 June 2004 and had the objectives of improving the accuracy of the web page classifier and the development of new classifiers for a number of other Internet data types. In the first phase specifications were defined quite generally while much greater specification was created for the second phase. In phase 1 the requirements consisted of a statement on system accuracy (50%) over 3 types of scam in a category scheme of 4 classes restricted to the analysis of web pages. The client provided a manually classified corpus of about 8000 documents. The delivery time was 6 months from project commencement. The project team consisted of 2 linguists, 1 computational linguist and 3 software engineers.

In Phase 2 the contract had more specification of data sources to be scrutinized, entity recognition and performance requirements, plus in each case retrieval

mechanisms had to be developed and for one source the corpus had to be compiled. The team was expanded with another linguist and computational linguist. Throughout the project there were 4 research assistants providing support roles, 3 PhD students making exploratory research contributions and two linguist advisors.

Results –Phase 1

The results of the first phase of the web page classifier for the scam classes as applied to an audit corpus are presented in figure 1. This corpus was unseen by the development team and made available by ASIC at the time of final delivery of the system. The processing was conducted by the ASIC staff and the project team was given one week in which to request revisions to ASIC's manual classifications. The Scamseek classifier in this instance also identified 4 scams that had been manually misclassified by ASIC. An audit report authored and approved by both parties was completed.

| ASIC Class | Computed Class | | |
|------------|----------------|----------|-------|
| | Scam | Non-scam | TOTAL |
| Scam | 18 | 26 | 44 |
| Non-Scam | 6 | 1525 | 1531 |

Figure 1. Performance results for an audit corpus of the phase 1 web page classifier conducted on 21/10/2003 in the ASIC offices.

The results in Figure 1 represent performance values of: Precision=.75, Recall=.41, and F-value=.53 and are to be contrasted with the laboratory results of the completed system on the training corpus using 10-fold cross validation of: Precision = .74, Recall = .35, F=.48. A baseline 1000 single words has F=.21. ASIC was entirely satisfied with these results and made a commitment to a larger project in Phase 2.

Results – Phase 2

| | Web Pages | Corpus 2 | Corpus 3 |
|----------------------|-----------|----------|------------|
| Precision | .744 | .850 | .852 |
| Recall | .528 | .834 | .639 |
| F-value | .618 | .844 | .730 |
| Scam /non-scam texts | 373/6391 | 686/1483 | 1395/13716 |

Figure 2. The performance results for web pages classifiers and 2 other classifiers for identifying scams as delivered to ASIC.

The results of the second phase of the web page classifier for the scam classes applied to the corpus are presented in figure 2. ASIC was satisfied by the performance of the system in phase 1 not to require a second audit corpus assessment. Figure 2 provides results for 3 separate corpora, web pages as in the phase 1 experiments, and two other corpora developed for phase 2. The Web Pages result represents results for the system delivered to ASIC as of 30 June, 2004. The exact nature of the other corpora cannot be presented due to security obligations. The performance figures are from 10-fold cross-validations.

Conclusions

The Scamseek project is a success for ASIC in that it is operational 24 hours a day 7 days a week. In its first operational run it discovered an activity that has

since been taken to the stage of prosecution. The estimate of savings in human effort in its monitoring role is the order of 100 to 1, as previously ASIC had to read 80 documents to find one of interest they now read 5 documents to find 4 of interest. The estimate in savings to the community by bringing speedier detection and intervention of scams cannot be estimated readily but is likely to be of the order of tens of millions of dollars. The project in both of its phases outperformed its contract specifications, and came in on time and within budget. ASIC is not able to release all details about the technology but has released the following summary statement: “The Scamseek technology is deployed in such a way that any scam proposal on any Internet channel that is generated in Australia or directed at Australians is highly likely to be scrutinised”.

The research contribution has been significant in that it is the first project that has used Systemic Functional Grammar for automated text classification. Solutions to serious problems in practical text classification, namely unbalanced classes, and the integration of semantic and unigram language models have also been developed.

The project has also made a significant contribution to the issues of software engineering in language technology as it has shown that computational linguistics research can be performed in the context of industrial objectives, and that good software engineering enhances the productivity of an experiment program.

ⁱ See ASIC Media Release 04-178: Grammax Investment Club operating unlicensed investment clubs.