# Extracting and Visualizing Trust Relationships from Online Auction Feedback Comments

**John O'Donovan** and **Barry Smyth**

Adaptive Information Cluster

School of Computer Science and Informatics

University College Dublin

Ireland

*firstname.lastname@ucd.ie*

**Vesile Evrim** and **Dennis McLeod**

Semantic Information Research Laboratory

Department of Computer Science

University of Southern California

Los Angeles

*lastname@usc.edu*

## Abstract

Buyers and sellers in online auctions are faced with the task of deciding who to entrust their business to based on a very limited amount of information. Current trust ratings on eBay average over 99% positive [13] and are presented as a single number on a user's profile. This paper presents a system capable of extracting valuable negative information from the wealth of feedback comments on eBay, computing *personalized* and *feature-based* trust and presenting this information graphically.

## 1 Introduction

In 2005, eBay Inc. had an annual growth rate of 42.5% and in February 2006 was receiving over three million user feedback comments every day. This growth reflects the increased access to online markets throughout the world. As a result of this rapid growth rate, users find themselves more frequently in the situation where they must trust a complete stranger to uphold their part of a business transaction. Traditional mechanisms for evaluating the trust of a potential buyer or seller do not apply in the online world, where in many cases the only information provided is a buyer or seller's username and there are orders of magnitude more potential transactors. [14] In this paper we address the problem of providing a reliable mechanism for trusting users in online auctions. We argue that the current trust system on eBay is massively biased towards positive comments and show how *AuctionRules*, a trust-extraction algorithm can be applied to the wealth of freetext comments to capture subtle indications of negativity, producing a more scaled range of trust based on the sentiment found in comments. To achieve this, the algorithm classifies freetext comments using a small list (<30 nouns) of salient features in the domain. Our evaluations show this algorithm outperforms seven popular classification algorithms by up to 21%.

Currently, eBay displays a non-personalized trust score for a potential transactor. We describe a trust propagation mechanism which uses output from *AuctionRules* to generate personalized trust between two potential transactors. For demonstration purposes, we present a replica interface to a standard auction site in which trust between specific users and per-feature trust can be computed by *AuctionRules* on the fly and visualized as a social network graph with edge length and thickness as functions of both strength and goodness of the trust relationship.

## 2 Background Research

Background for this paper is in three main areas. Firstly, we will examine related research in the area of trust and reputation in online environments. The core algorithm in this paper uses techniques borrowed from the field of natural language processing (NLP). Next we discuss relevant work in this area of NLP. Lastly, a brief examination of research on constructing and visualising social networks is presented.

### 2.1 Trust and Reputation on the Web

A large amount of recent research has examined the varying notions of trust and reputation on the web. [7] [10] [13] Marsh's work in [7] goes some way towards formalising trust in a computational sense, taking into account both it's social and technological aspects, defining "context-specific interpersonal trust" as the trust one individual must place in another with respect to a specific situation. This is the notion of trust examined in this paper. Previous work by the authors in [10] examines trust and reputation in the context of recommender systems, wherein a user's history of contributions to the recommendation process was examined and used to weight future contributions.

Resnick highlights some relevant points which affect the current eBay reputation system in [13] and [14]. Buyers reputation matter less since they hold the goods until they are paid. Feedback can be affected by the person who makes the first comment, ie: feedback can be reciprocated. Retaliatory feedback and potential for lawsuits are strong disincentives for leaving negative comments. Anonymity is possible in eBay since real names are not revealed and the only thing validated at registration is an email address. Users can choose not to display feedback comments. "Unpaid item" buyers cannot leave feedback [1], and users can agree to mutually withdraw feedback [1]. Furthermore eBay inc admit that negative feedback is removed from newer users of the system [1] All of these points help to explain the lack of negative comments on eBay. Of course this does not mean that customers are satisfied. The eBay forums [1] highlight the fact that false advertising does occur on eBay. This should lead to more negative comments, but they are not being displayed. Xiong et al. [15] provide further reasoning for the imbalance of positive comments on eBay. Our proposed model of trust for eBay should provide a more realistic scale than the existing system.

---

[1] http://forums.ebay.com

## 2.2 Classifying Natural Language Comments

For many years research effort has been dedicated to the use of lexical techniques to classify free responses in areas ranging from online auctions [5] to movies [11] and for correcting students short-answer exams. [2]. Processing of freetext comments requires four main components, morphology (prefixes and suffixes), syntax checking, deciding on a semantic meaning and finally, taking some action based on that meaning. In this paper we are especially interested in the classification of comments with *negative* sentiment. Yukari et al. [6] proposed a method to extract potentially unsatisfied customers by applying text mining to data from a customer satisfaction survey. This work introduced the idea of salient *satisfaction factors* as a mechanism for classifying negative comments. Later in this paper a *feature-based* trust metric is introduced, which was inspired by work in [6]. Pang et al. attempt to classify freetext documents according to *sentiment* in [11], highlighting bad performance of machine learning algorithms such as naïve bayes and support vector machines in the task of sentiment classification. Pang et al. conclude that classification of free text according to sentiment is a challenging problem. Gamon et al address the same sentiment classification problem in [3], using data from car reviews. Gamon et al. also applied machine learning techniques to this task and found poor performance. They developed a tailor-made clustering algorithm which uses a small amount of domain knowledge in the form of a *stop-list* and *go-list* of features known to be salient in the domain. The *AuctionRules* algorithm described later in this paper defines small negative and positive lists in a similar manner to that in [3].

Gamon et al. [3] introduce the concept of varying the level of granularity during the classification procedure, finding that "varying the level of granularity of analysis allows the discovery of new information". The *AuctionRules* algorithm can record finer grained information during classification, to present extra information to users.

Hijikata et al. [5] focus on summarization of freetext comments in online auctions, proposing a technique for analysing the underlying social networks in online auctions and using this information to present a summary of comments to users. In this work, the social network for their test data is defined by a crawler algorithm which takes all the feedback comments written by a buyer who wrote a comment for a target seller. As a result of this data crawling technique it appears that the data collected does not represent a natural social network, and if large portions of freetext data come from single individuals there is a potential for overfitting during computation. A web application is presented in [5] where users can see comment summaries represented as a set of bar charts for easy analysis.

## 2.3 Construction and Visualisation of Social Network Graphs

As people conduct larger proportions of their daily business on the web, the ability to analyse the resultant social networks is becoming more important. Technologies such as FOAF (Friend-Of-A-Friend) are becoming more popular in blogging communities, allowing users to directly specify their social networks. Visualisation tools such as *PieSpy* [9] (used later in this paper) allow for dynamic display and fast assessment of complex graphs. Mitton [9] assesses a range of social networks mined from IRC channels. Golbeck uses social networks to compute trust scores and shows resulting visual-

isations derived from the *FilmTrust* movie ratings website in [4].

## 3 Extracting Trust From Feedback Comments

To address the problem of unnaturally high trust ratings on eBay, we look to the freetext comments and apply a classification algorithm tailored for capturing subtle indications of negativity in those comments. The situation arises frequently where users are afraid to leave a negative comment for fear of retaliatory feedback comments which could damage their own reputation. [13]. In many of these cases, a positive feedback rating is made, but the commenter still voices some grievance in the freetext comment. This is the type of subtle but important information the *AuctionRules* algorithm attempts to extract.

### 3.1 The *AuctionRules* Algorithm

*AuctionRules* is a classification algorithm with the goal of correctly classifying online auction comments into positive or negative according to a threshold. *AuctionRules* capitalizes on the restrictive nature of online markets: there are a limited number of salient factors that a user (buyer or seller) is concerned about. This is reflected in feedback comments. We define a set of seven core *feature sets* for which the algorithm will compute granular trust scores. The following sets have a coverage of 62% of the comments in our database. The algorithm can obtain semantic information from 62% of the comments at a fine grained level. It is shown in our experimental analysis how we can maintain over 90% coverage using this algorithm. The terms in brackets are contents of each feature set.

- *Item* - The quality/condition of the product being bought or sold. (*item, product*)
- *Person* - The person the user makes the transaction with. (*buyer, seller, eBayer*, *dealer*)
- *Cost* - Cost of item, cost of shipping, hidden costs etc. (*expense, cost*)
- *Shipping* - Delivery of the item, security, time etc. (*delivery, shipping*)
- *Response* - Communication with the other party, emails, feedback comment responses. (*response, comment, email, communication*)
- *Packaging* - The packaging quality/condition of the item (*packaging*)
- *Payment* - how the payment will be made to the seller, or back to buyer for return (*payment*)
- *Transaction* - the overall transaction quality (*service, transaction, business*)

This technique enables us not only to compute a personal trust score between individual users, but also to provide more granular information on a potential transactor. For example: "User x is very trustworthy when it comes to payment, but shipping has been unsatisfactory in the past", This granular or *contextual* trust draws on the wealth of information in comments and can uncover hidden problems which the current trust system on eBay might overlook.

Figure 1 details this algorithm working on a sample comment which had a positive rating on eBay. Each term in the comment and up to four preceding terms are passed into an
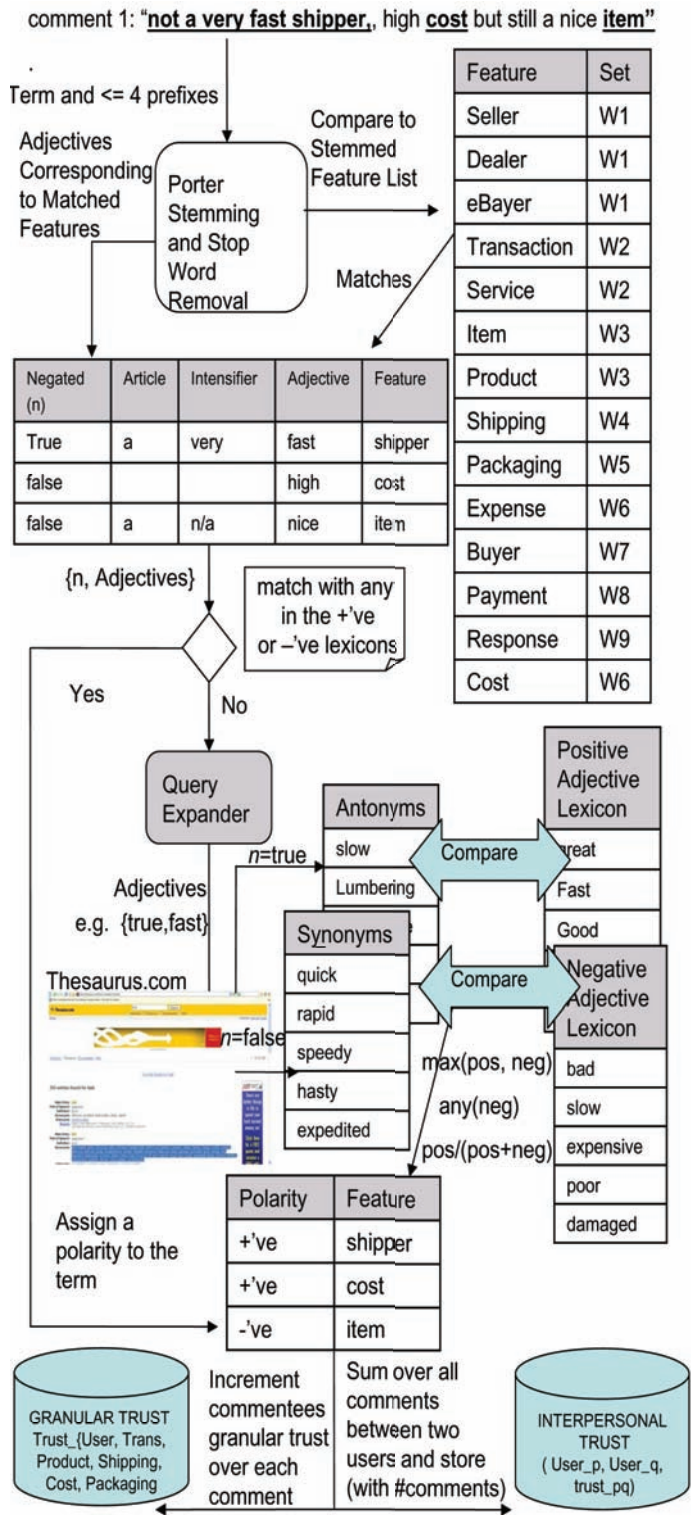
Figure 1: The *AuctionRules* comment classification algorithm.

implementation of the Porter stemming algorithm [12]. The standard porter stemmer uses many rules for removing suffixes. For example, all of the terms in $c$ are conflated to the root term *"connect"*. $c$ = *connect*, *connected*, *connecting*, *connection*, *connections*. This reduces the number of terms and therefore the complexity of the data. The stemmer algorithm was modified to also stem characters not used by *AuctionRules*, such as "?, !, *, (, )" for example.

Data dimension is reduced by removal of stop-words. Google's stop-word list[2] was used as the removal key. A future addition to the algorithm would put an automatic spelling corrector at this point. There were 11 different spelling occurrences of the word 'packaging' for example. Each stemmed term is compared against the stemmed terms from the feature list. If a match is found the algorithm examines the preceding term types. This is shown graphically in Figure 1. The algorithm recognises five term types:

- *nouns* the words contained in the feature sets.
- *adjectives* (eg: 'nice', 'good') from a web list
- *intensifiers* (e.g 'very', 'more') list of 20.
- *articles* 'a' or 'the'
- *negators* (e.g. not, 'anything but') from a stored list.

The table on the left in Figure 1 shows the order in which these terms are analysed. From the five terms, two can provide semantic meaning about the feature: adjectives and negators. If an adjective is found without a negator, it is compared to an arrays 20 positive and an array of 20 negative adjectives. If a negator is found as the second or third preceding term, the process works with positive and negative arrays switched. If a match is found, the polarity for that feature is recorded.

If no match is found, *AuctionRules* uses query expansion, by calling an interface to an online thesaurus which returns an array of 20 synonyms. If a negator is present, the interface returns an array of 20 antonyms, and these lists are compared in a similar manner to our short lexicon of positive and negative adjectives. The matching results are recorded in three ways: a) $max(pos, neg)$ b) $any(neg)$ and c) $\frac{neg}{pos+neg}$. In the case of (c) the polarity is recorded according to a configurable threshold $\alpha$. Two separate trust databases are maintained: *granular* or *contextual* trust which is the trust computed for each feature for a user over all of the comments analysed. Equation 1 shows contextual trust $t$ as a multi valued set of trust scores associated with each feature. Here, $f$ denotes a particular feature and $t_f n$ is the associated trust score. The second trust database is *interpersonal* trust which is the average trust value recorded for all features on every comment made between two users.

Of course not every comment will contain words from the feature lists above. In cases where no features are found, the algorithm performs a straightforward count of positive and negative terms using the query expansion module where necessary. In this manner, coverage is maintained at over 90%, and many of the unclassifiable comments are foreign language or spelling mistakes.

### 3.2 Personalizing Trust by Propagation

Interpersonal trust is given by Equation 2. This is a personalised trust score between two users. If users have been di-
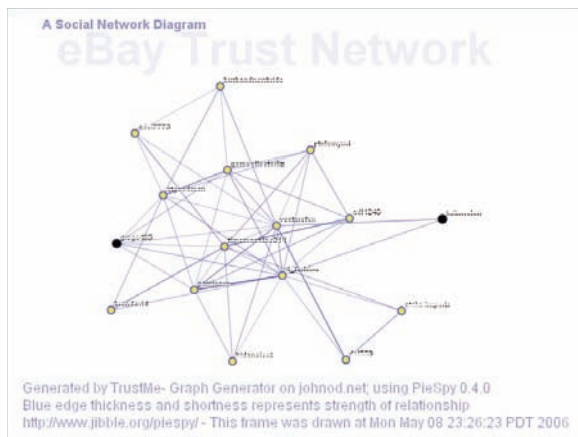
Figure 2: Visualizing the trust-network generated by *AuctionRules*. (usernames have been obfuscated at the request of eBay inc.)

rectly involved in a transaction, this is simply the classifiers value of their transactions (a, b, or c above). Equation 2 shows how these trust scores can be computed along a path connecting the two users. In Equation 2 $\oplus$ represents some combination of the scores at each node in the path between the target and source users. Currently this can be any of the following four techniques below. The problem of combining values in the trust graph is a research in itself. This will be dealt with in more detail in a future work.

- *weightedDistance* - The average trust score over all the edges in the shortest path, discounted by the distance from the source.

- *meanPath* - The average trust score over all the edges in the shortest path between the source node and target node.

- *twoPathMean* - The average of the *meanPath* of the shortest path in the graph and the *meanPath* of the second shortest path.

- *SHMPath* - The simple harmonic mean of the trust scores over all edges in the shortest path.

$$t_{granular}\varepsilon\{t_{f1}, t_{f2}, ...t_{fn}\} \qquad (1)$$

$$t_{i,n} = t_{i,j} \oplus t_{j,k} \oplus ... \oplus t_{m,n} \qquad (2)$$

### 3.3 Visualisation of Trust Information

Before introducing our presentation strategy, we define two important metrics, firstly a scalar trust *value* which is represents trust either between two users, or for one user on a particular feature. Secondly, we define trust *strength*. This is based on the number of transactions between two users in the case of interpersonal trust, or as the number of comments used to calculate the feature trust score in the case of contextual trust.

To generate on-the-fly visualisations of the resultant trust network, we modified a version of *PieSpy* an open source graphing tool [9] by Mutton for visualising social networks. Our interpersonal trust database is a set of triples of the form: $e_{(i,j)} = (user_i, user_j, t_{(i,j)})$, where $t_{(i,j)}$ comes from Equation 2. This was translated into XML and fed to the graphing tool. Modifications were made to the existing code to highlight target and source users in the graph. Figure 2 shows the
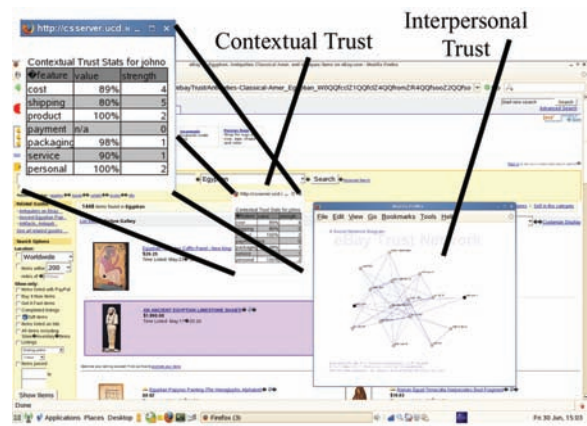


Figure 3: Integration of the trust system into an online auction. (usernames have been obfuscated at the request of eBay inc.)

visualisation output from a demonstration set of users in our database. (this graph has much higher than average interconnectedness.) The graphing tool caters for trust *value* and trust *strength* by representing value as the thickness of an edge and strength of the trust-relationship as the length of a line. In the graph, shorter lines mean a higher number of comments were used in the classification. The demonstration system in Figure 3 shows the trust graph popup when a mouseOver occurs on a username. To visualise contextual trust, a popup table with feature names, associated trust scores and strengths is used. This is shown as the smaller popup in Figure 3.

## 4 Experimental Evaluation

We examine four factors in our evaluations. The *accuracy* of the *AuctionRules* classifier with respect to other techniques from machine learning. We examine accuracy from a Mean Absolute Error perspective and by creating confusion matrices to examine performance with respect to false negative and positive classifications. As the system uses a very small amount of domain knowledge, and a limited number of features, we must examine the *coverage* of *AuctionRules* over our comments database. Finally we make a comparison between the *scale* of trust achieved by *AuctionRules* against the current eBay scale.

### 4.1 Data

Figure 4 explains the environment in which we test our algorithm. Data is crawled from the online auction site according to the crawler algorithm below. Importantly, unlike it's machine learning counterparts, *AuctionRules* requires *no* knowledge of the comment set it has to classify. The feature lists used by the algorithm are generic and should work on any set of online auction comments.

**Algorithm 4.1:** CRAWL($String\ url\_list,\ int\ maxbound$)

```
while n < maxbound
  for i ← 1 to #Items_on_page
    followSellerLink();
    for j ← 1 to #Comments_on_page
      db.add(bId, sId, comment, sTrust, bTrust);
    n ← n + 1;
return ;
```
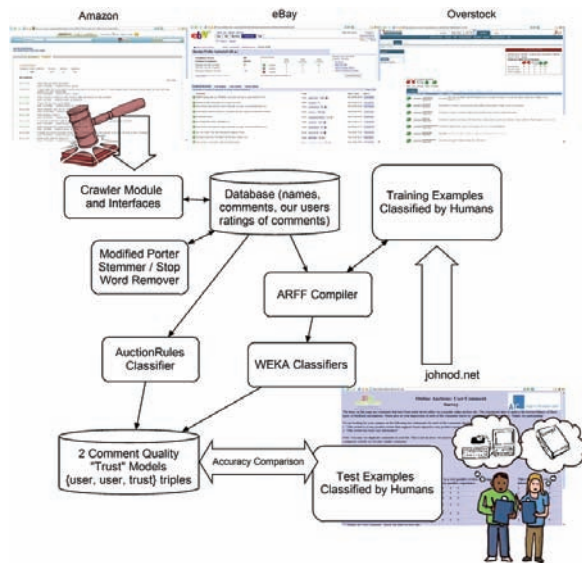
Figure 4: Graphical overview of the trust-modelling process. (Current implementation only uses eBay as a source for ratings.)

10,000 user comments were collected from the auction site using the crawler. As a social network visualisation was planned, data was collected from within a domain with highly interconnected and contained repeat-purchase users. After a manual analysis of a range of sub-domains of the auction site, Egyptian antiques was chosen as the data domain as it appeared to meet the prerequisites to a reasonable degree. Although a large number of comments were collected, only 1000 were used in our experimental analysis.

**User-Provided Training Data**

In order to test any comment classification algorithm a benchmark set was required. 1000 comments were compiled into an online survey[3] and rated by real users. In this survey, users were asked to rate the positiveness of each comment on a Likert scale of 1 to 5. 10 comments were presented to a user in each session. Each comment was made by different buyers about one seller. Users were required to answer the following:

- How positive is the comment (Average rating: 3.8442)

- How informative is the comment (Average rating: 3.1377)

- Would you buy from this seller (Average rating: 4.0819)

Currently only results from the first question are used to develop and test *AuctionRules*. For future experiments we may incorporate results from the other questions. Permission was sought from eBay inc. to use the information from the eBay website in our experiments.

## 4.2 Comparing Accuracy Against Machine Learning Techniques

To examine classification accuracy of *AuctionRules*, it was tested against 7 popular algorithms. We chose three rule-based learners, *Zero-r*, *One-r*, *Decision Table*, one tree

---
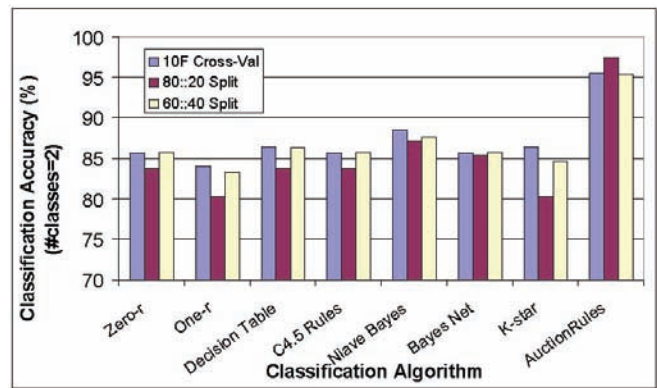
[3]www.johnod.net/Surveyone.jsp



Figure 5: Classification Accuracy [classification distribution from user evaluations: 36% positive, 63% negative, using a threshold of 4 or higher for a positive comment.]

learner *C4.5 rules*, two Bayes learners, *Naive Bayes* and *BayesNet* and a lazy learning algorithm *K-Star*.

Figure 5 shows results of this experiment. For each algorithm we performed three runs. a 60:40 train-test split, an 80:20 split, and a 10-fold cross validation of the training set, which randomly selects a training set from the data over 10 runs of the classifier and averages the result. In the experiment each algorithm made a prediction for every value in the test set, and this prediction was compared against the training set. *AuctionRules* beat all of the other classifiers in *every* test we performed, achieving over 90% accuracy in all of the evaluations, 97.5% in the 80:20 test, beating the worst performer *K-Star* by 17.5%, (relative 21.2%) and it's closest competitor *Naive Bayes* by 10.5%, giving a relative accuracy increase of 12.7%.

In addition to numerical accuracy, we examined where the high accuracy results were coming from more closely by assessing the confusion matrix output by the algorithms. This was necessary since prediction of false negatives would have an adverse effect on the resulting trust graph. This phenomenon has been discussed by Massa in [8] with respect to the *Moleskiing* application, and Golbeck in [4] with respect to the TrustMail application. Table 4.2 shows *AuctionRules* outperforming all of the other algorithms by predicting no false negatives. This is taken as a good result since in a propagation computation negative values should contain more weight because of their infrequency. When a value is presented to a user directly however, false positives are more damaging for an auction environment. *AuctionRules* also performs very well for false positives with a rate of 4.5%, half that of the closest competitor *One-r*. All of the algorithms displayed similar trend to the ones in Table 4.2, which shows results of the 80:20 classification experiment which had a test set of 234 comments. It was found during accuracy evaluations that there was a strong correlation between the number of feature terms recognised and the final accuracy of the classification. For our coverage experiments, we addressed the number of *hits* found with respect to coverage. This is detailed in the following section.

## 4.3 Coverage and Distribution Experiments

To assess the coverage of the *AuctionRules* feature-based trust calculator we examined the number of feature term hits

| | AuctionRules | | NaiveBayes | | Decision Table | | One-r | |
|---|---|---|---|---|---|---|---|---|
| | +'ve | -'ve | +'ve | -'ve | +'ve | -'ve | +'ve | -'ve |
| +'ve | 91.4 | 0 | 84.1 | 1.2 | 84.6 | 1.2 | 77.3 | 8.1 |
| -'ve | 4.7 | 4.7 | 11.1 | 2.9 | 12.3 | 1.7 | 8.5 | 5.9 |

Table 1: Confusion matrices showing percentage true and false negatives and positives for four of the algorithms tested. [All of the other algorithms had similar results to the ones displayed.]
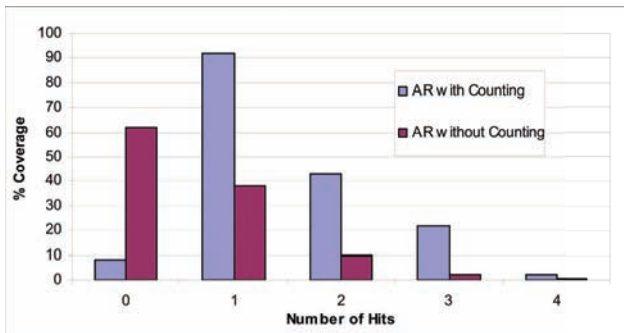


Figure 6: Comparison of the coverage and hit ratio in *AuctionRules*

that occur during classification. Coverage was tested in two modes. Firstly, the standard feature-based mode. In this case, 62% of the 1000 comments tested contained at least one hit from the feature list. This provides us with semantic knowledge about a large number of comments. However there is a sharp reduction in coverage when we look for comments with more than one hit. To increase coverage *AuctionRules* uses simple term counting to supplement its feature-based approach in cases where it fails to find any terms. When counting is applied the coverage is greatly increased with over 90% of the comments getting at least one hit. After manual examination, it is clear that majority of the other 8% can be attributed to misspellings, foreign language and other noise.

A distribution analysis was performed between the *AuctionRules* generated trust values and the current and found that the new trust values do capture some unseen negativity from feedback comments. This is manifested in a slightly more scaled distribution relative to the current eBay values. From 600 comments that were analysed in the distribution experiment, 90 comments which had been rated as 93 to 100 percent positive were taken out of that bracket by *AuctionRules*. This may seem like a small amount, but compared with the current system which has less than 1% negative ratings, *AuctionRules* produced 16%.

## 5 Conclusions and Future Work

The main contribution of this paper is an algorithm for extracting *personalized* and *contextual* trust from the wealth of freetext comments on online auction sites. The algorithm operates on the assumption that online auction transactions can be categorized into a relatively small set of features. *AuctionRules* can extract context-specific and personal trust both

retroactively and on the fly as new comments are added. There are a range of uses for the extracted trust scores. In this paper we have shown one such use in a pop-up visualisation of the resulting trust network for a demonstration subset of the eBay marketplace. This visualisation also shows per-feature trust in a pop-up window for a buyer or seller. In our evaluations we show that even using a small lexicon of key features, coverage is still maintained above 90%. We show that the *AuctionRules* classifier beats seven popular learning algorithms at this classification task by up to 21% in accuracy tests using very minimal domain knowledge. *AuctionRules* has a favourable false negative rate for the classifier of 0% compared to up to 6% from benchmark algorithms and a false positive rate of 4.7% compared with 8.5% to 12.3% for the benchmark algorithms.

## 6 Acknowledgements

## References

[1] Brian Burke. Ebay town-hall meeting. Transcript, Senior Manager, eBay Marketplace Rules and Policy, eBay Inc, San Jose, CA. Feb 28th. 2006. available at http://pics.ebay.com/aw/pics/commdev/TownHallTrans cript_022806.pdf.

[2] Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu, and Martin Chodorow. Enriching automated essay scoring using discourse marking. In Manfred Stede, Leo Wanner, and Eduard Hovy, editors, *Discourse Relations and Discourse Markers: Proceedings of the Conference*, pages 15–21. Association for Computational Linguistics, Somerset, New Jersey, 1998.

[3] Michael Gamon, Anthony Aue, Simon Corston-Oliver, and Eric K. Ringger. Pulse: Mining customer opinions from free text. In *IDA*, pages 121–132, 2005.

[4] Jennifer Ann Golbeck. *Computing and applying trust in web-based social networks*. PhD thesis, University of Maryland at College Park, College Park, MD, USA, 2005. Chair-James Hendler.

[5] Yoshinori Hijikata, Hanako Ohno, Yukitaka Kusumura, and Shogo Nishida. Social summarization of text feedback for online auctions and interactive presentation of the summary. In *IUI '06: Proceedings of the 11th international conference on Intelligent user interfaces*, pages 242–249, New York, NY, USA, 2006. ACM Press.

[6] Yukari Iseyama, Satoru Takahashi, and Kazuhiko Tsuda. A study of knowledge extraction from free text data in customer satisfaction survey. In *KES*, pages 509–515, 2004.

[7] S. Marsh. *Formalising Trust as a Computational Concept*. PhD thesis, Department of Mathematics and Computer Science, University of Stirling, 1994.

[8] Paolo Massa and Bobby Bhattacharjee. Using trust in recommender systems: an experimental analysis. *Proceedings of 2nd International Conference on Trust Managment, Oxford, England*, pages 221–235, 2004.

[9] Paul Mutton. Inferring and visualizing social networks on Internet Relay Chat. In *Eighth International Conference on Information Visualization (IV04)*, page 9 total. IEEE, July 2004.

[10] John O'Donovan and Barry Smyth. Trust in recommender systems. In *IUI '05: Proceedings of the 10th International Conference on Intelligent User Interfaces*, pages 167–174. ACM Press, 2005.

[11] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, 2002.

[12] M. F. Porter. An algorithm for suffix stripping. *Readings in information retrieval*, pages 313–316, 1997.

[13] Paul Resnick and Richard Zeckhauser. Trust among strangers in internet transactions: Empirical analysis of ebay's reputation system. *The Economics of the Internet and E-Commerce. Volume 11 of Advances in Applied Microeconomics.*, December 2002.

[14] Paul Resnick, Richard Zeckhauser, Eric Friedman, and Ko Kuwabara. Reputation systems. *Communications of the ACM*, December 2000. Vol. 43, No. 12.

[15] Li Xiong and Ling Liu. Building trust in decentralized peer-to-peer electronic communities. In *Fifth International Conference on Electronic Commerce Research (ICECR-5)*, 2002.