Association for Information Systems
AIS Electronic Library (AISeL)

MCIS 2016 Proceedings

Mediterranean Conference on Information Systems (MCIS)

2016

# Challenges in Short Text Classification: The Case of Online Auction Disclosure

Yichen Li
*University of Auckland*, yli781@aucklanduni.ac.nz

Arvind Tripathi
*University of Auckland*, a.tripathi@auckland.ac.nz

Ananth Srinivasan
*University of Auckland*, a.srinivasan@auckland.ac.nz

# CHALLENGES IN SHORT TEXT CLASSIFICATION: THE CASE OF ONLINE AUCTION DISCLOSURE

*Completed Research*

Yichen Li, University of Auckland, New Zealand, yli781@aucklanduni.ac.nz

Arvind Tripathi, University of Auckland, New Zealand, a.tripathi@auckland.ac.nz

Ananth Srinivasan, University of Auckland, New Zealand, a.srinivasan@auckland.ac.nz

## Abstract

*Text classification is an important research problem in many fields. We examine a special case of textual content namely, short text. Examples of short text appear in a number of contexts such as online reviews, chat messages, twitter feeds, etc. In this research, we examine short text for the purpose of classification in internet auctions. The "ask seller a question" forum of a large horizontal intermediary auction platform is used to conduct this research. We describe our approach to classification by examining various solution methods to the problem. The unsupervised K-Medoids clustering algorithm provides useful but limited insights into keywords extraction while the supervised Naïve Bayes algorithm successfully achieves on average, around 65% classification accuracy. We then present a score assigning approach to this issue which outperforms the other two methods. Finally, we discuss how our approach to short text classification can be used to analyse the effectiveness of internet auctions.*

*Keywords: Short Text Classification, Online Auctions, Text analytics*

## 1 Introduction

In spite of the considerable work in the area of numerical or categorical data analysis, several challenges remain in the processing of textual data (Losiewicz, Oard, & Kostoff, 2000). The past decades have witnessed extensive theoretical and empirical studies in both unsupervised and supervised learning methods in the fields of machine learning and text analytics. Among them, text document clustering is the application of statistical cluster analysis as well as data mining techniques to unstructured digital text documents (Steinbach, Karypis, & Kumar, 2000) while document classification is defined as the effort to determine how a document should be categorized under a given heading (Borko & Bernick, 1963). Currently automatic document classification is widely used in spam filtering (Cormack & Lynam, 2007), subject categorization (Chuang & Chien, 2003), authorship identification (Zheng, Li, Chen, & Huang, 2006), and sentiment analysis (Moraes, Valiati, & Neto, 2013) to name a few areas of enquiry. The importance of applying text mining methods in the information systems discipline has been recently highlighted by Debortoli, et al (2016).

Recently, increasing attention has been paid to classifying short texts with the rapid popularization of e-commerce and online communication. Short texts exist in numerous contexts, such as instant chat messages, twitter feeds, online product reviews, feedback mechanisms, news comments, intra-transactional Q&A (Basu, et al., 2015) and so on. Researchers tend to perform text categorization and include the predicted categories in a set of subsequent statistical analyses to explore relationships among the categories and other outcome variables. The subsequent analysis is sensitive to mis-classification; thus quality categorization is imperative for effective modelling that relies on the categorization results.

Due to its unique characteristics, short text classification is deemed to be demanding and challenging. Firstly, a short text is sparse data which is highly reliant on context. It is problematic to select powerful language features since shared context and word co-occurrences are insufficient for using valid distance measures. Secondly, short text always appears in large quantity, resulting in conventional document

classification running into problems such as labelling bottlenecks. It is intractable to assign labels manually in a sizable training set while limited tagged instances might not be adequate for machine learning. Thirdly, prior efforts in the area of automatic coding of short text failed to guarantee satisfactory accuracy because much of content is only partially related to the coding task (Larkey & Croft, 1995). Another difficulty is the common existence of non-standard terms and noise such as misspellings, grammatical errors, abbreviations, slang words or even foul language. An appropriate method should be tolerant of a certain degree of such "anomalies". Consequently, how to reasonably represent and choose salient, invariant and discriminatory features (Forsyth & Holmes, 1996), effectively reduce spatial dimensionality and noise, and make the best use of those limited hand-labelled instances are stimulating questions for short text classification.

In this research, our focus is on a particular case of short text occurrences, the online auction Q&A, which can be extremely short, cryptic, sparse, and ungrammatical. Moreover, it always fails to provide sufficient term occurrences and it is often complicated to identify underlying sentiment information. This study compares two very popular methodologies; the first is the unsupervised K-Medoids clustering approach, which is deemed to be effective in document clustering, and the second is the classical Naive Bayes algorithm, which is an example of supervised learning. We then address the shortcomings of these two approaches for our context and develop a score assigning (n-gram) approach that shows superior performance. Finally, we demonstrate how the categorization can be used for modelling and understanding various phenomena that characterize online auctions.

The rest of the paper is organized as follows: the next section presents a review of related literature on text clustering and classification. This is followed by brief introductions of data collection process and the adopted internet auction platform. Subsequently, the applications of K-Medoids algorithm and Naïve Bayes classifier to the data are explained. This is followed by a description of the score assigning approach and the applicability of the results for the purpose of modelling.

# 2 Literature Review

While tremendous achievements have been obtained in numerical or categorical data analysis, the processing of textual data still remains to be perfected (Losiewicz et al., 2000). In general, machine learning algorithms are divided into two approaches, namely, unsupervised learning and supervised learning. Unsupervised learning seeks to uncover hidden features in a set of unlabelled data. The groups are unknown beforehand and the objective of document clustering is to determine the category of each observation based on an appropriate distance measure which both maximizes between-group variation as well as minimizes within-group variation. No training data is available to evaluate whether the classification is accurate or not. However in supervised learning, group labels are defined beforehand, and the categorization process is executed by adopting a learned text classifier which works with a training set of human annotated examples. This literature review provides an overview of text document clustering as well as text document classification, which are typical implementations of unsupervised and supervised learning respectively.

## 2.1 Text Document Clustering

Text document clustering is the application of statistical cluster analysis to textual data such that documents with homogeneous meanings and connotations are assigned to the same clusters (Neto, Santos, Kaestner, Alexandre, & Santos, 2000). A large strand of literature has focused on design and performance of the clustering algorithms. Hierarchical and partitioning are two major families in this line of work (Steinbach et al., 2000). The concept of the partitioning clustering approach is to represent clusters using several central vectors, which need not be elements of their corresponding clusters. Observations

are iteratively grouped into k non-overlapping clusters in which each observation belongs to the proximate collection with the nearest mean (Jain, 2010). Therefore, this kind of centroid-based clustering is also called K-Means-Type clustering which converts a partitioning process into an optimization problem. Notwithstanding superior performance in practice, one of the main handicaps of this approach is that the value k is an input parameter and needs to be predetermined (Milligan & Cooper, 1985). Hierarchical clustering, also known as connectivity-based clustering, always produces a dendogram which portrays an extensive hierarchy of groups that combine with each other by use of a proper metric (Willett, 1988). Different clusters are formed according to different distances and dissimilarities, which are marked and presented along the y-axis of a dendogram.

Text document clustering has long been a widely discussed problem in the field of information retrieval and processing. Szymanski (2011) presented an approach to successfully automate the categorization of Wikipedia search outcomes by implementing a combination of several clustering algorithms (Szymański, 2011). Though the method behaved well in segregating Polish Wikipedia articles, its effectiveness on other databases failed to be adequately shown. The partitioning K-Means approach is adopted to approximate and select journal papers with the help of prearranged patterns. Although the performance was acceptable due to the decrease in search time, the semi-automatic process restricted the search and learning capability and efficiency to a knowledge database. In similar vein, Ma and Xu (2012) intended to design a decision support system to assist government and other private research funding agencies to automatically select candidate research projects into corresponding subject areas utilizing ontology-based text clustering technologies (Ma et al., 2012).

## 2.2 Text Classification

Document classification is defined as the effort to determine how a document should be categorized under a given heading (Borko & Bernick, 1963). This effort seeks to investigate how automatically and efficiently determine the category according to certain attributes or a set of given rules. In spite of the effort involved in building and maintaining rules, supervised text classification is of considerable interest in the realm of machine learning since accuracy can be surprisingly high provided features are cautiously selected and refined by experts. In this section, we delineate the importance as well as several applications of feature selection and discuss previous investigations on classification of short text.

### 2.2.1 Feature Selection

With an increasing amount of textual data generated, proliferated and stored over the internet, high-performance information retrieval is demanding and time-consuming without proper simplification and organization of the content. Efficient feature selection is one solution to this problem. Automatic feature selection involves the process of filtering redundant attributes based on corpus statistics and collecting relevant terms to construct predictive models using higher-level orthogonal dimensions (Yang & Pedersen, 1997). The feature set can be seen as the most informative and indicative subset of the training set. It improves and expedites a certain classifier by reducing the size of powerful vocabulary. In addition, it mitigates the risks of overfitting problems and strengthens the categorization performance by accounting for noisy attributes.

In tasks of text classification, feature selection plays an indispensable role in training the text classifiers. Among all those techniques, K-Nearest Neighbors (KNN), Naïve Bayes and Support Vector Machine (SVM) are the most widely acknowledged and utilized. Zhang and Lee (2003) implemented bag-of-words as well as bag-of-n-grams feature models and a tree kernel function to explore solutions to automatic question classification. They noted that SVM turned out to be the best performer which is capable of exploiting underlying syntactic information of questions (Zhang & Lee, 2003). Apart from employing different metrics individually, hybrid approaches may yield unforeseen outcome development. Rogati

et al. (2002) also conducted an empirical study to compare the comprehensive performance of five predominant feature selection metrics and explored whether the synergy of certain criteria would promise higher efficiency or effectiveness using four leading text classifiers, namely, Naïve Bayes, KNN, SVM and Rocchio-style classifier. Consequently, the chi-square metric coupled with document frequency and information gain criteria have witnessed increments in performance concerning the four classifiers, not to mention the removal of sparse words (Rogati & Yang, 2002). Furthermore, the incorporation of several methods have been acclaimed to be conducive to optimize feature selection process by resolving dependency and redundancy problems. Das (2001) proposed a boosting-based hybrid algorithm which assimilated part of the competitive edges of a Wrapper method into filter methods (Das, 2001). This method turned out to be better performing than its original components. However, one of the limitations of this combinational algorithm is that it demands substantial computations for pairwise correlations, which is intractable to scale to a huge dataset. Apart from the hybrid of feature selection methods, Jeong et al. (2016) have proposed a promising framework leveraging the outstanding feature weighting capability of text summarization and the categorization ability of text classification. The result of the experiment indicated that the combination can successfully improve the performance of the individuals.

### 2.2.2 Short Text Classification

Much of the literature has focused on long text contexts. These successful results now need to be investigated in short text contexts for their applicability and potential improvement. One example of this work is presented by Sun (2012) using a straightforward but scalable method which achieved favourable categorization accuracy (Sun, 2012). The approach started with a manual extraction of representative word combinations. A Term Frequency and Inverse Document Frequency (TFIDF) weighting scheme were adopted to ensure topic-specificity of the query word sequences that can represent as much content as possible. Then it searched for a limited set of hand-coded texts that are most relevant to the query instances and determined the category heading according to the highest score and vote based on previous search results. In contrast, Li and Qu (2013) subscribed to the view that the classical TFIDF weighting factor is not effective for short text classification. They argued that even the refined TFITF algorithm (ITC) which substitutes term frequency with its logarithmic form has conspicuous imperfections on account of the high dependence on the quality of training collection (Li & Qu, 2013). Hence, the authors demonstrated a solution which overcomes the deficiencies to a large extent. The new hybrid functions amalgamated the Document Distribution Entropy algorithm as well as the Position Distribution Weight algorithm together, and it outperformed the conventional methods.

In terms of handling grammatical errors in a document, it is argued that the n-gram-frequency-based classification method is ideal for text documents that come from noisy sources (Cavnar, 1995), where N-gram here refers to N contiguous sequence of letters. Considering that the occurrence rate of any spelling, grammatical or machine recognition errors tend to be relatively low and that every single string is broken down into small pieces, the negative effect of any textual errors is tolerable since only a negligible part of the whole document is affected. One limitation of this method is that the importance of an N-gram profile only depends on its occurrence frequency, ignoring statistics for less-frequent N-grams, which might be informative in some circumstances (Cavnar & Trenkle, 1994).

## 3 Conceptual Framework

The influences of pre-configuration and information disclosure including pre-transactional passive voluntary disclosure, post-transactional feedback and intra-transactional live interaction on online auction outcome have been extensively discussed in previous literature. Undoubtedly, all those disclosures serve the goal of alleviating adverse selection issues and building trust among auction participants. It is not likely that potential buyers will make bidding decisions as soon as they open a listing site; normally an

item will be listed for a fixed period of time thus people can consider, compare and then make up their minds. Voluntary item descriptions and previous feedback rating or qualitative reviews are deemed to be conventional channels for buyers to collect information. Potential buyers may ask more questions regarding aspects that are not mentioned in the

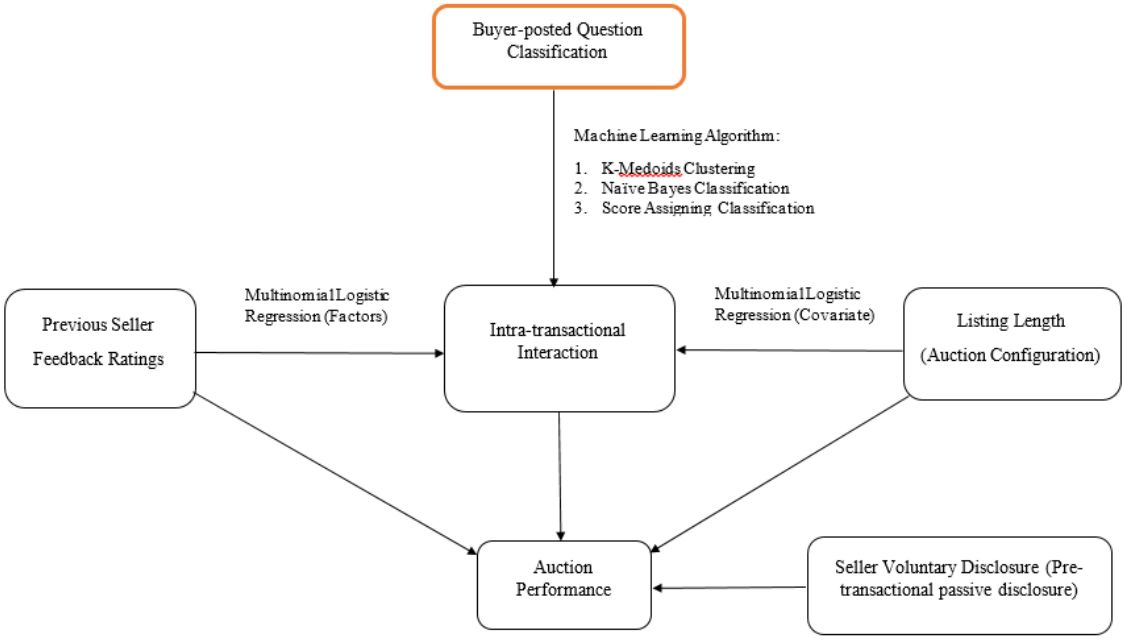


*Figure 1. Conceptual Model*

product description and explore further explanations concerning information that has already been revealed by the seller. This work obviously acknowledges the significant impact that online word-of-mouth exerts on auction activity. However, due to the fact that intra-transactional information largely consists of textual data, quantification is necessary for further analyses. The conversion from qualitative data to quantitative data is undertaken by classifying the buyer-initiated questions according to given rules. Consequently, the qualitative features of intra-transactional interaction during an auction and the quantification of textual auction Q&A are considered along with the role of different genres of information disclosure and their impacts on auction outcomes in the comprehensive conceptual framework (Figure 1).

# 4 Methodology

## 4.1 The auction database and data collection

We examine auction listing data from a widely used online auction platform. A feature of this platform is that it allows interaction among buyers and sellers while an auction is in progress through a Q&A facility. To control for exogenous effects on our results, we restrict our attention to the trading of used cars. This is a product class with sufficient ambiguity to generate a sufficient level of buyer-seller interaction through the Q&A mechanism. Compared to new cars, there exists a higher level of information asymmetry that requires clarification beyond information that is voluntarily disclosed as part of the listing. Such information is typically unique to the particular used car on offer resulting in the exchange.

We collected data that consisted of 14812 listings over a 8-month period of time. This data set represents 42940 instances of short text requiring classification.

## 4.2 Q&A as an example of short text

In a second-hand goods market, sellers publish various kinds of information and buyers tend to be passive receivers (Dellarocas, 2003). However with the evolution of online auction mechanisms, real-time interaction has drawn greater attention in the past few years and it enabled buyers to seek more information according to their own needs. Nowadays major online auction websites have all introduced online Q&A forums where buyers can post questions to a specific seller concerning a particular listing and sellers are required to respond. This interactive mechanism establishes relationships "between any pair of question issuers and answer providers in the public space" (Wang & Chiang, 2009) with a view toward improving efficiency in the market mechanism.

The purpose of this study is to compare methods for categorizing short texts. The online auction Q&A, which can be extremely short, cryptic, sparse, and ungrammatical, is a unique type of short text that proliferated in predominant C2C online trading platforms. The "ask seller a question" forum of a large horizontal intermediary auction platform is used to conduct this research. Potential bidders are allowed to ask for information or make comments on whatever they want regarding a listing. These are easy to access at the bottom of a listing. This forum provides an exclusive bi-directional platform that not only highlights the live timing of interaction, but also assists the progress of benefiting the public by providing people valuable information before they cast their bids.

Reliable classifications of internet auction Q&A can be corner stones to further understand customers' participations and responses in online trading communities. For instance, researchers may explore how different kinds of buyer-posted questions affect the performance of auctions. The quality of such exploration is reliant on high quality classification of short text instances.

## 4.3 Application of K-Medoids method

We started to categorize the text by applying the K-Medoids clustering algorithm in the realm of unsupervised learning. Being considered as an amalgamation and extension of the K-Means standard algorithm as well as the "Medoidshift" algorithm, the K-Medoids approach is acclaimed for its robustness to noise and sensitivity to outliers since it selects k representative data points as centroids (or Medoids) rather than a mean point. Considering the observations in the entire dataset, it effectively minimizes a sum of absolute distances between the points and the chosen centre instead of a sum of squared distances .

The most successful application of the K-Medoids algorithm is the Partitioning Around Medoids (PAM) algorithm which works with the dissimilarity matrix. A fast heuristic technique is utilized aiming to spot an acceptable solution efficiently. In addition, the PAM algorithm also generates vivid visual representations, the clusplot as well as the silhouette plot. The clusplot depicts the objects with their correlations and simultaneously portrays the relative range, shape as well as locations of clusters, while the silhouette plot pictures how well each data point is positioned within its cluster. Moreover, the corresponding silhouette width provides a way to evaluate the optimal numbers of clusters, which helps to address a big problem in automatic clustering. Since the category of the object is unknown in clustering and there are no training examples to examine whether the categorization is correct or not, it remains to be a big challenge to determine the number of clusters. Usually the approach is faced with a trade-off: on the one hand, having fewer clusters makes categories easy to understand and discriminable, on the other hand, having more clusters allows to identify more significant segments and more subtle differences between segments.

There have been plenty of empirical attempts to provide support for the efficient and effective use of K-Medoids algorithm on document classification, which aims at identifying natural groups of a set of documents according to a given rule; for instance, the dissimilarity measure. In this research, we also attempted to apply this algorithm to cluster questions and comments that buyers posted during online auction. We chose the pam() function in R software to apply the algorithm, which is able to suggest an optimal number of clusters based on Silhouette width obtained through rounds of trials. A stemming process is implemented to identify words with a common meaning and form as being identical, which is broadly used in text analytics. The system chose 40 as the optimal number of clusters and listed the keywords in each cluster.

While this provided some insights into the effectiveness of classification, we identified several shortcomings with this approach. Several reasons account for the failure. Firstly, there are always empty clusters in the output and the number of observations in empty clusters remains relatively high. The crucial factor is that we use high-frequency words to do the clustering. The high frequency vocabulary is identified and extracted from the whole corpus which is a collection of all the "documents". But the content in each and every document is not rich enough normally containing only one or two sentences. Since a large number of documents don't include these clusters at all, these documents cannot be extracted by using high-frequency words. Consequently they are clustered into a group which has no common words to represent the whole cluster. The second reason is due to an inherent limitation of unsupervised learning techniques. A major challenge in unsupervised text classification lies in its inability to reveal latent traits of textual data and therefore appropriately interpret the outputs under context-free circumstances. Therefore, despite the fact that the result presents several features, it is burdensome to assign meaningful labels to those clusters based on a single word or sequence. Further, we are hampered in our ability to ensure one question only belongs to one dominant cluster since some text may express two or more latent categories at a time.

## 4.4 Application of Naïve Bayes method

### 4.4.1 Predetermined Clusters

Classification involves developing pre-determined clusters. This belongs to the supervised learning where group labels are defined as the first step. Discriminatory and meaningful predetermined categories are an essential prerequisite for classification, whose importance has always been overlooked since an overwhelming majority of problems are limited to binary classifications. Additionally, for some difficult problems with more than two categories, the clusters are determined based on common sense. Nevertheless, for problems that require a new set of categories, reasonable determination and explicit clarification are imperative. In this section, we pre-determined the categorization clusters. We then examine the performance of the application of the methods to our dataset. The question categories are listed in Table 1.

| No. | Cluster Name | Description |
|---|---|---|
| 1 | Product Description and Quality Inspection | Questions and comments regarding asking more/detailed information of used car auction listing and the intention of looking or inspecting the vehicle. Posts in this category tend to concentrate on vehicle's previous usage and maintenance. Depending on the current condition evident in the listing, customers may request more visual information to make better decisions. |

| | | |
|---|---|---|
| 2 | Seller information and credibility | Questions and comments asking for more information of the sellers such as contact number or email address to access the sellers' characteristic and furthermore mitigate the seller uncertainty. |
| 3 | Transaction and shipment | Questions and comments which involve discussions on what kind of transaction method they are going for and corresponding shipment details if they win the bid. |
| 4 | Negotiation | Questions and comments concerning price negotiation and swapping enquiries. |
| 5 | General questions and comments | This category includes all the general questions and comments which cannot be classified into any other clusters above. To ensure the integrity of this research, all the questions in the dataset are included in these five clusters, regardless whether the post makes sense. |

*Table 1. General descriptions of five pre-determined clusters for Naïve Bayes classification*

### 4.4.2 Data Decryption and Results

Using a data set of 600 randomly selected questions, we divided the data into two parts, the first 550 made up the training set while the remaining 50 comprised the hold-out set. The program automatically adopts the trained classifier to categorize elements in the hold-out set, then compares the machine-generated results with the human-labelled ones so as to calculate the accuracy. Every time we run the module, the dataset is randomly shuffled, thus a different set of training data is generated to train a new classifier with divergent feature constitutions. Again, this classifier is utilized to cluster the rest of rows, and the performance measure can be computed later. Fig. 2 below shows a typical output of the program. We can see that, apart from accuracy percent, it also returns a list of most informative features which has been used in the classifier trained in the last run. For instance, a buyer-initiated question or comment which mentions the feature "number" or "why" is approximately 42 or 14 times more likely to be in the second group (seller information and credibility) than in fourth group (negotiation), which is quite reasonable and justifiable since the second cluster represents people's concern that the seller will act opportunistically. They are more likely to ask for more information of the seller such as phone number and physical address or doubt why the seller wants to sell the car or why the price is extremely low.

```
Naive Bayes Algorithm Accuracy Percent: 69.38775510204081 %
Most Informative Features
                  number = True                2 : 4      =     42.3 : 1.0
                    pick = True                3 : 4      =     27.0 : 1.0
                 address = True                2 : 1      =     17.5 : 1.0
                    cash = True                3 : 2      =     17.2 : 1.0
                 listing = True                4 : 1      =     17.0 : 1.0
                    view = True                2 : 4      =     16.7 : 1.0
                    away = True                3 : 1      =     15.3 : 1.0
                 arrange = True                3 : 1      =     15.3 : 1.0
                     pay = True                3 : 1      =     15.3 : 1.0
                 bidding = True                2 : 1      =     14.3 : 1.0
                     why = True                2 : 4      =     14.1 : 1.0
                   keen = True                 4 : 1      =     13.7 : 1.0
                   from = True                 3 : 4      =     12.3 : 1.0
                  price = True                 4 : 1      =     12.2 : 1.0
                   take = True                 3 : 1      =     11.8 : 1.0
Naive Bayes algorithm ended
```

*Figure 2.* *An example of Naïve Bayes classifier output*

In order to estimate how the performance of the Naive Bayes classifier will generalize to other independent classification tasks, a general accuracy percent should be calculated through a model validation technique. In this study, the cross validation method is adopted for two reasons. Firstly, this method is easily implemented in the Python programming language. Secondly, as reported in previous research, cross-validation is acclaimed to be adequate for testing hypotheses which are based on data already observed rather than new unknown data, particularly when further observations are expensive, time-consuming or inaccessible (Kohavi, 1995). We conducted 12 rounds of cross-validation with different partitions. The graph (Fig. 3) below depicts the accuracy rate recorded in every round. The results fluctuated over the 12 rounds, reaching a peak of somewhere around 75% and a bottom at roughly 55%. We can conclude that the accuracy rate of Naive Bayes classifier on buyer-initiated questions and comments classification is around 65% on average.

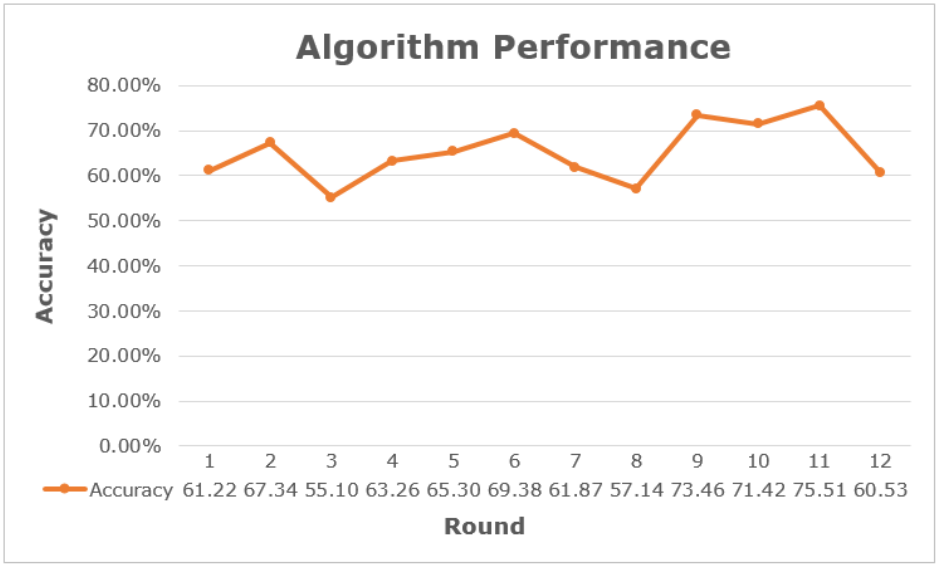


*Figure 3* *Accuracy rate recorded in 12 rounds of cross-validation*

## 4.5 CLASSIFICATION - SCORE ASSIGNING APPROACH

In this section, we propose a new classification algorithm that performs well in the context of our data set. In a nutshell, this algorithm combines a semi-automated feature selection process and a fully-automated text classification according to a set of scores assigned to extracted features (clue words and sequences). Unlike the previous algorithms, no human-labelled training set is needed, instead, all the features are assigned to different predetermined clusters beforehand.

The algorithm starts with clue words and sequence extraction. The sequence here is defined as two or more consecutive single words. It has been shown in prior literature that word combinations are more likely than single words to have discriminating power in information retrieval, especially the words that constitute the sequence. However, single words should not be overlooked since they still can be informative given that buyer-initiated questions are often short, cryptic and ungrammatical. Therefore, features which are extracted to carry out classification contain both single words as well as sequences. The feature selection is implemented by automatic sequence tokenization and manual labeling; then we obtain a pool of clue words and sequences. Note that a single word and a word combination are both regarded as one feature. After that, we check how many times each feature occurs in a large data pool in order to provide a basis for score assigning. Normally, group labels are defined beforehand, and the categorization process is executed by adopting a learned text classifier which works with a training set of human-annotated examples. As its name suggests, pre-determined groups are necessary for classification. This method also adopts the clusters that we pre-defined beforehand. But different from conventional classifier which needs an appropriate manually labeled training set, this algorithm categorizes the features rather than original texts. Specifically, a set of scores are assigned to each clue word or sequence to indicate how important that feature is in terms of determining the cluster that a piece of text should go to. Ultimately, the automatic text classification process is undertaken by a Java program. The whole algorithm including detailed score assigning rules and machine-based classification are explained and amplified in the next section.

Three sets of data are employed in this classification methodology. The first one includes 1000 randomly chosen buyer-initiated questions and comments, which are used to identify and extract clue words and sequences. The second one contains 30000 elements which are prepared for frequency checking. Note that there is no need to pre-process or manually label these two datasets with regard to pre-determined clusters, all we want is to keep the original data. The third dataset is identical to the dataset we use in Naïve Bayes classification; 600 buyer-initiated questions are randomly selected and human-labelled with categories. The manually labeled dataset will be the testing set that is utilized to evaluate how the algorithm performs.

To assess the performance of our short text classification algorithm, the 600 randomly selected questions and comments are analyzed. A coder who was not aware of the purpose of this study was invited to conduct the manual labeling. The outcome of the human labeling is sorted by the categorization number in ascending order and utilized as a reference. In order to compare algorithm performances with the Naïve Bayes method, we use the same dataset in the method I as our testing set. Subsequently, the algorithm is applied to automatically classify those 600 questions and comments. Then the two sets of results are put together for assessment. We use a score sheet to evaluate the performance of automatic clustering, add one penalty score every time we find an inappropriately grouped text, then accumulate the penalty scores regarding each cluster.

Let P1, P2, P3, P4 and P5 denote the number of incorrect memberships (penalty scores) of each cluster, let C1, C2, C3, C4 and C5 denote the counts of selected samples in each cluster, and the performance is calculated by comparing the automatic and manual clustering:

$$Performance = 1 - \frac{\sum_{i=1}^{5} \frac{P_i}{C_i}}{5} * 100\%$$

The Performance evaluation sheet below (Table 2) illustrates how the performance is evaluated.

| Cluster | Counts (C) | Penalty Score (P) | Error rate |
|---|---|---|---|
| 1 | 239 | 26 | 0.1088 |
| 2 | 51 | 6 | 0.1176 |
| 3 | 27 | 4 | 0.1481 |
| 4 | 188 | 32 | 0.1702 |
| 5 | 95 | 18 | 0.1895 |
| Average Error Rate | | 0.1468 | |
| Performance | | 0.8532 | |

*Table 2. Performance Evaluation*

The computational classification program exhibits high flexibility, repeatability and extensibility. This approach does not need a training set, the classification is carried out according to categorized features in advance and corresponding scores assigned to them. Consequently, classification accuracy can be enhanced by optimizing the memberships in the clue words and sequences list as well as the score assigning rules. The results show that this approach works well in the domain that we investigating. It outperform more classical approaches of unsupervised and supervised learning methods.

# 5 Using Classification Results

Accurate classification allows us to investigate relationships of the nature outlined in the model in Figure 1. For example, we might wish to investigate how seller feedback count and listing length drive the formation of buyer-initiated questions. Thus the dependent variable should be a categorical variable which is set to represent the different genres of questions or comments that buyers post online. Since online Q&A are qualitative data, a classification process is necessary to convert textual information into the categorical form for subsequent analysis (in this case we propose the use of a technique such as multinomial logistic regression). Such analysis is often vulnerable and sensitive to misclassification, thus quality categorization is imperative for doing this effectively. A compelling and high-quality categorization for intra-transactional buyer-initiated Q&A is of paramount importance to accurately measure and distinguish buyers' intentions of raising a question or leaving a comment. Using our classification approach which successfully achieved an average accuracy of 85.67%, we categorized the buyer-initiated questions into the five predetermined clusters. In a nutshell, this algorithm combines a semi-automated feature selection process and a fully-automated text classification according to a set of scores assigned to extracted features (clue words and sequences).

As per the relationships outlined in the conceptual model shown in Figure 1 a number of opportunities for further analysis arise. For example, analyses of how numerical feedback ratings and listing duration motivate people's choices of raising intra-transactional questions can be carried out using multinomial logistic regression. This method allows for a categorical dependent variable with more than two classes and it is capable of checking interactions among independent variables. In this research, seller reputation ratings (rated as "high", "medium", or "low") and buyer-initiated question types are both nominal; they are utilized as our dependent variable and one of the independent variables respectively, listing length

is regarded as a metric covariate. By choosing one of the outcome variable categories as the baseline comparison group (base), we can run five models with different categories treated as the reference level.

.

# 6 Conclusions

While this study is a significant step towards our understanding of short text classification, there are a number of issues that could motivate future work in this area. Our data comes from a specific context, intra-transactional disclosure during online auctions and therefore, we don't know how our proposed algorithm will perform on short text from different contexts. This work could be extended to look at other product categories beyond used cars that forms the basis of the current work. Variation in the information content of different product categories could shed light on the efficacy of our approach. Following our observation from the previous section, the results of the multinomial regression analysis could be useful in terms of understanding clearly the nature of the impact of questions and answers under varying contingency situations.

In this research we investigated the phenomenon of online auctions with the overall objective of understanding some fundamental characteristics of how these mechanisms work. In particular we acknowledge the vast amount of work that has appeared in the literature that has examined such things as seller reputation, quality information provided and transaction details that have resulted in increasing the market efficiency of online auctions. In our work we focused on a particular type of information disclosure: specifically information exchange while an auction is in progress. Such information appears in the form of short text and it is imperative that we classify this textual data accurately for use in model based investigations. We show that a score assigning n-gram based approach for classification outperforms more traditional unsupervised and supervised learning approached of classification. Finally we demonstrate how these accurate classifications can be used to investigate the determinants of various short text categories. We believe that a proper understanding of the role of such interaction among buyers and sellers in online auctions will lead to better design of platforms and therefore contribute to improving the quality of overall outcomes for all stakeholders in the mechanism.

# Reference


Basu, A., Lee, Youngjin, Srinivasan, A., and Tripathi, A. (2015). An Empirical Analysis of Intra-Transactioanl Disclosure in Internet Auctions. *Proceedings of Twelfth European Mediterrenean & Middle Eastern Conference in Information Systems (EMCIS)*.

Borko, H., & Bernick, M. (1963). Automatic document classification. *Journal of the ACM (JACM), 10*(2), 151-162.

Cavnar, W. (1995). Using an n-gram-based document representation with a vector processing retrieval model. *NIST SPECIAL PUBLICATION SP*, 269-269.

Cavnar, W. B., & Trenkle, J. M. (1994). N-gram-based text categorization. *Ann Arbor MI, 48113*(2), 161-175.

Chuang, S.-L., & Chien, L.-F. (2003). Enriching web taxonomies through subject categorization of query terms from search engine logs. *Decision support systems, 35*(1), 113-127.

Cormack, G. V., & Lynam, T. R. (2007). Online supervised spam filter evaluation. *ACM Transactions on Information Systems (TOIS), 25*(3), 11.

Das, S. (2001). *Filters, wrappers and a boosting-based hybrid for feature selection.* Paper presented at the ICML.

Debortoli, S., Muller, O., Junglas, I., and vom Brocke, H. (2016) Text mining for Information Systems Researchers: An Annotated Topic Modeling Tutorial. Communications of the AIS, Vol. 39, Article 7, Available at http://aisel.aisnet.org/cais/vol39/iss1/7.
Forsyth, R. S., & Holmes, D. I. (1996). Feature-finding for text classification. *Literary and Linguistic Computing, 11*(4), 163-174.
Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters, 31*(8), 651-666.
Jeong, H., Ko, Y., & Seo, J. (2016). How to Improve Text Summarization and Classification by Mutual Cooperation on an Integrated Framework. *Expert Systems with Applications, 60, 222-233.*
Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection.* Paper presented at the Ijcai.
Larkey, L. S., & Croft, W. B. (1995). Automatic assignment of icd9 codes to discharge summaries. *University of Massachusetts*.
Li, L., & Qu, S. (2013). Short Text Classification Based on Improved ITC. *Journal of Computer and Communications, 2013*
Losiewicz, P., Oard, D. W., & Kostoff, R. N. (2000). Textual data mining to support science and technology management. *Journal of Intelligent Information Systems, 15*(2), 99-119.
Ma, J., Xu, W., Sun, Y.-h., Turban, E., Wang, S., & Liu, O. (2012). An ontology-based text-mining method to cluster proposals for research project selection. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 42*(3), 784-790.
Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*(2), 159-179.
Moraes, R., Valiati, J. F., & Neto, W. P. G. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications, 40*(2), 621-633.
Neto, J. L., Santos, A. D., Kaestner, C. A., Alexandre, N., & Santos, D. (2000). Document clustering and text summarization.
Steinbach, M., Karypis, G., & Kumar, V. (2000). *A comparison of document clustering techniques.* Paper presented at the KDD workshop on text mining.
Sun, M. (2012). How does the variance of product ratings matter? *Management Science, 58*(4), 696-707.
Szymański, J. (2011). Self–Organizing Map representation for clustering Wikipedia search results *Intelligent Information and Database Systems* (pp. 140-149): Springer.
Wang, J.-C., & Chiang, M.-J. (2009). Social interaction and continuance intention in online auctions: A social capital perspective. *Decision support systems, 47*(4), 466-476.
Willett, P. (1988). Recent trends in hierarchic document clustering: a critical review. *Information Processing & Management, 24*(5), 577-597.
Yang, Y., & Pedersen, J. O. (1997). *A comparative study on feature selection in text categorization.* Paper presented at the ICML.
Zhang, D., & Lee, W. S. (2003). *Question classification using support vector machines.* Paper presented at the Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval.
Zhang, J. (2006). The roles of players and reputation: evidence from eBay online auctions. *Decision support systems, 42*(3), 1800-1818.