Questionnaire Free Text Summarisation Using Hierarchical Classification

Matias Garcia-Constantino, Frans Coenen, P-J Noble and Alan Radford

Abstract This paper presents an investigation into the summarisation of the free text element of questionnaire data using hierarchical text classification. The process makes the assumption that text summarisation can be achieved using a classification approach whereby several class labels can be associated with documents which then constitute the summarisation. A hierarchical classification approach is suggested which offers the advantage that different levels of classification can be used and the summarisation customised according to which branch of the tree the current document is located. The approach is evaluated using free text from questionnaires used in the SAVSNET (Small Animal Veterinary Surveillance Network) project. The results demonstrate the viability of using hierarchical classification to generate free text summaries.

1 Introduction

The proliferation and constant generation of questionnaire data has resulted in substantial amounts of data for which traditional analysis techniques are becoming harder and harder to apply in an efficient manner. This is particularly the case with respect to the free text element that is typically

Matias Garcia-Constantino
Department of Computer Science, The University of Liverpool, Liverpool, L69 3BX, UK,
e-mail: mattgc@liverpool.ac.uk

Frans Coenen
Department of Computer Science, The University of Liverpool, Liverpool, L69 3BX, UK,
e-mail: coenen@liverpool.ac.uk

P-J Noble
School of Veterinary Science, University of Liverpool, Leahurst, Neston, CH64 7TE, UK,
e-mail: rtnorle@liverpool.ac.uk

Alan Radford
School of Veterinary Science, University of Liverpool, Leahurst, Neston, CH64 7TE, UK,
e-mail: alanrad@liverpool.ac.uk

included in questionnaire surveys. A technique used to understand and extract meaning from such free text is text summarisation. The motivations for text summarisation vary according to the field of study and the nature of the application. However, it is very clear that in all cases what is being pursued is the extraction of the main ideas of the original text, and the consequent presentation of these ideas to some audience in a coherent and reduced form. Many text summarisation techniques have been reported in the literature; these have been categorised in many ways according to the field of study or to other factors inherent to the text. Jones *et al.* [16] proposed a categorisation dependent on: the input that is received, the purpose of the summarisation and the output desired. An alternative categorisation is to divide the techniques according to whether they adopt either a statistical or a linguistic approach.

Text classification, in its simplest form, is concerned with the assignation of one or more predefined categories to text documents according to their content [8]. The survey presented by Sebastiani [27] indicates that using machine learning techniques for automated text classification has more advantages than approaches that rely on domain experts to manually generate a classifier. As in the case of more established tabular data mining techniques, text classification techniques can be categorised according to whether they attach a single label (class) to each document or multiple labels. The approach proposed in this paper to generate text summaries is based on the concept of hierarchical text classification, which is a form of text classification that involves the use of class labels arranged in a tree structure. Hierarchical text classification is thus a form of multi-label classification. As Sun and Lim state [29], hierarchical classification allows a large classification problem to be addressed using a "divide-and-conquer" approach. It has been widely investigated and used as an alternative to standard text classification methods, also known as *flat classification* methods, in which class labels are considered independently from one another.

Although they are intended for different forms of application, text summarisation and text classification share a common purpose, namely to derive meaning from free text (either by producing a summary or by assigning labels). The reason why text summarisation can be conceived of as a form of text classification is that the classes assigned to text documents can be viewed as an indication (summarisation) of the main ideas, of the original free text, in a coherent and reduced form. Coherent because class names that are typically used to label text documents tend to represent a synthesis of the topic with which the document is concerned. It is acknowledged that a summary of this form is not as complete or as extensive as what many observers might consider to be a summary; but, if we assign multiple labels to each document then this comes nearer to what might be traditionally viewed as a summary. However, for anything but the simplest form of summarisation, the number of required classes will be substantial, to the extent that the use of flat classification techniques will no longer be viable, even if a number of

such techniques are used in sequence. A hierarchical form of classification is therefore suggested. By arranging the potential class labels into a hierarchy multiple class labels can still be attached to documents in a more effective way than if flat classifiers were used. The effect is to permit an increase in the number of classes that can be used in the classification. Our proposed hierarchical approach uses single-label classifiers at each level in the hierarchy, although different classifiers may exist at the same level but in different branches in the hierarchy.

The advantages of using the proposed hierarchical text classification for text summarisation are as follows: (i) humans are used to the concept of defining things in a hierarchical manner, thus summaries will be produced in an intuitive manner, (ii) hierarchies are a good way of encapsulating knowledge, in the sense that each node that represents a class in the hierarchy has a specific meaning or significance associated with it with respect to the summarisation task, (iii) classification/summarisation can be achieved efficiently without having to consider all class labels for each unseen record, and (iv) it results in a more effective form of classification/summarisation because it supports the incorporation of specialised classifiers, at specific nodes in the hierarchy.

The rest of this paper is organised as follows. Related work is briefly reviewed in Section 2, and a formal definition of the proposed free text summarisation mechanism is presented in Section 3. Section 4 gives an overview of the SAVSNET (Small Animal Veterinary Surveillance Network) project questionnaire data used for evaluation purposes with respect to the work described in this paper. Section 5 describes the operation of the proposed approach. A comprehensive evaluation of the proposed approach, using the SAVSNET questionnaire data, is presented in Section 6. Finally, a summary of the main findings and some conclusions are presented in Section 7.

2 Related work

The text summarisation techniques proposed in the literature take into account the field of study and factors inherent to the text to be summarised. Afantenos *et al.* [1] provided what is referred to as a "fine grained" categorisation of text summarisation factors founded on the work of Jones *et al.* [16], and formulated the summarisation task in terms of input, purpose and output factors. The input factors considered were: (i) the number of documents used (single-document or multi-document), (ii) the data format in which the documents are presented (text or multimedia) and (iii) the language or languages in which the text was written (monolingual, multilingual and cross-lingual). The purpose factors were sub-divided according to: (i) the nature of the text (indicative or informative), (ii) how specific the summary must be for the intended audience (generic or user oriented) and (iii) how specific the summary must be in terms of the domain or field of study (general purpose or domain specific). The most significant output factors (amongst others) were sub-divided according to whether the summary needed to be: (i) complete,

(ii) accurate and (iii) coherent. The "traditional" phases of text summarisation that most researchers follow are identified in [2], namely: (i) analysis of the input text, (ii) transformation of the input text into a form to which the adopted text summarisation technique can be applied and (iii) synthesis of the output from phase two to produce the desired summaries.

To the best knowledge of the authors, the generation of text summaries using text classification techniques has not been widely investigated. Celikyilmaz and Hakkani-Tür [3] presented an "automated semi-supervised extractive summarisation" approach which used latent concept classification to identify hidden concepts in documents and to add them to the produced summaries. Previous work by the authors directed at text summarisation can be found in [10] and [11]. In [10] a summarisation classification technique, called Classifier Generation Using Secondary Data (CGUSD), was presented, it was directed at producing text summarisation classifiers where there was insufficient data to support the generation of classifiers from primary data. The technique was founded on the idea of generating a classifier for the purpose of text summarisation by using an alternative source of free text data and then applying it to the primary data. In [11], a semi-automated approach to building text summarisation classifiers, called SARSET (Semi-Automated Rule Summarisation Extraction Tool) was described, however this required substantial user intervention.

In [26], the integration of text summarisation and text classification is more synergic. Saravanan *et al.* proposed an approach to compose a summariser and a classifier integrated within a framework for cleaning and preprocessing data. They make the point that composition is invertible, meaning that summarisation can be applied first to increase the performance of the classifier or the other way around. As Saravanan *et al.* indicate, the use of classification improves the generation of summaries with respect to domain-specific documents. In [15], text classification is used to classify and select the best extracted sentences from text documents in order to generate summaries. In [14], a system is proposed that identifies the important topics in large document sets and generates a summary comprised of extracts related to identified topics.

Unlike the approach presented in this paper, the aforementioned approaches all use flat classification techniques to achieve free text summarisation. Hierarchical text classification makes use of the hierarchical relationships within an overall class structure to "boost" the effectiveness of text classification. The idea of using hierarchies for text classification can be effectively extended and customized for specific problems that involve the hierarchical representation of document sets. Typically, a hierarchy of a corpus of text documents is represented either as a decision tree or as a Directed Acyclic Graph (DAG). There are three main models in hierarchical classification: (i) big-bang, (ii) top-down and (iii) bottom-up. The big-bang model uses a single classifier to assign one or more class labels from the hierarchy to each document. The top-down and bottom-up models are based on the

construction and application of classifiers in each level of the hierarchy where each classifier acts as a flat classifier within that level/branch. In the case of the top-down model, the taxonomy is traversed from the higher to the lower levels. In the bottom-up approach the taxonomy is traversed in the reverse manner to that adopted in the top-down model. The model adopted with respect to the work described in this paper is the top-down model.

There is a considerable amount of research that has been carried out concerning the top-down model using different classification techniques, such as: Support Vector Machine (SVM) [6, 19, 29], classification trees [21], path semantic vectors [9], Hierarchical Mixture Models [30], TF-IDF classification [4], k-nearest neighbour techniques [7], a variation of the Maximum Margin Markov Network framework [24], the Passive-Aggressive (PA) algorithm with latent concepts [22], neural networks [25], word clustering combined with Naive Bayes and SVM [5], multiple Bayesian classifiers [18] and boosting methods (BoosTexter and Centroid-Boosting) combined with Rocchio and SVM classifiers [12]. A comprehensive survey on hierarchical classification is presented in [28].

As was mentioned in the previous section, there are many advantages of using hierarchical text classification for text summarisation, advantages that indicate that the approach proposed in this paper is a viable alternative over existing text summarisation techniques. The main advantages over other text summarisation techniques are: (i) a more intuitive way of understanding the contents of a document because humans are used to the concept of defining things in a hierarchical way and (ii) the ability to handle large document sets due to the hierarchical approach's inherent divide-and-conquer strategy. It can be argued that a summary generated using this approach will be very similar to that generated using a multi-class flat classification or to systems that automatically assign tags to suggest topics [17]. However, our approach differs from these other techniques in that the resulting classes generated using the hierarchical text classification process are not isolated concepts. On the contrary they are related to each other due to their hierarchical nature, giving the domain expert a more coherent and clear insight of what a given document is about.

3 Proposed approach

The input to the proposed text summarisation hierarchical classifier generator is a "training set" of n free text documents, $D = \{d_1, d_2, \ldots, d_n\}$, where each document d_i has a sequence of m summarisation class labels, $S = \{s_1, s_2, \ldots, s_m\}$, such that there is a one-to-one correspondence between each summarisation label s_i and some class g_j. Thus the summarisation labels are drawn from a set of n classes $G = \{g_1, g_2, \ldots, g_n\}$ where each class g_j in G has a set of k summarisation labels associated with it $g_i = \{c_{j_1}, c_{j_2}, \ldots c_{j_k}\}$. The desired class hierarchy H then comprises a set of nodes arranged into p levels, $L = \{l_1, l_2, \ldots, l_p\}$, such as the one shown in Figure 1. Except at the leaf nodes each node in the hierarchy has a classifier associated with it.

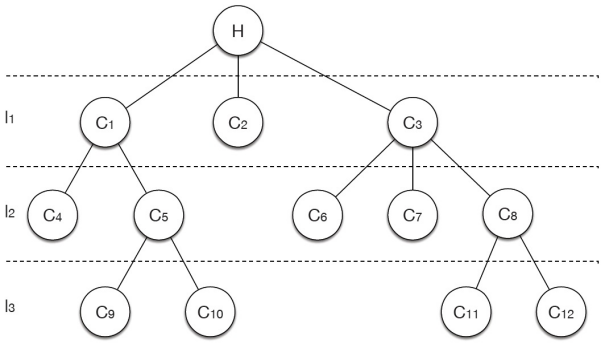

Fig. 1: Hierarchy of classes.

Since we are using a top-down model, the classifiers can be arranged according to two approaches in terms of the scope and dependency between levels: (i) cascading and (ii) non-cascading. In the cascading case, the output of the classifier at the parent nodes influences the classification conducted at the child nodes at the next level of the hierarchy (we say that the classification process "cascades" downwards). Thus a classifier is generated for each child node (except the leaf nodes) depending on the resulting classification from the parent node. The classification process continues in this manner until there are no more nodes to be developed. In the case of the non-cascading model each classifier is generated independently from that of the parent node.

In addition, two types of hierarchies are identified regarding the parent-child node relationship: single and multi-parent. The top-down strategy can be applied in both cases because, given a piece of text to be summarised, only one best child node (class) is selected per level. Examples of single and multi-parent hierarchies are shown in Figures 2 and 3.

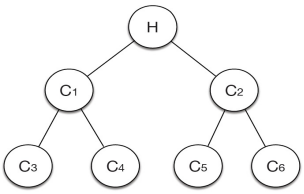

Fig. 2: Single-parent hierarchy

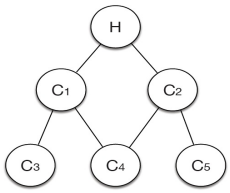

Fig. 3: Multi-parent hierarchy

The classifier generation process is closely linked to the structure of the hierarchy. When generating a cascading hierarchy we start by generating a classifier founded on the entire training set. For its immediate child branches we only used that part of the training set associated with the class represented by each child node. For non-cascading we use the entire training set for all nodes (except the leaf nodes), but with different labels associated with the training records according to the level we are at. The classifier generation process is described in more detail in Section 5.

Once the generation process is complete, the text summarisation classifier is ready for application to new data. New documents will be classified

by traversing the hierarchical classifier in a similar manner to that used in the context of a decision tree. That is, at each level the process will be directed towards a particular branch in the hierarchy according to the current classification. The length of the produced summary will depend on the number of levels traversed within the hierarchy.

4 Motivational example

The collection of questionnaires used to evaluate our approach was generated as part of the SAVSNET project [23], which is currently in progress within the Small Animal Teaching Hospital at the University of Liverpool. The objective of SAVSNET is to provide information on the frequency of occurrence of small animal diseases (mainly in dogs and cats). The project is partly supported by Vet Solutions, a software company whose software is used by some 20% of the veterinary practices located across the UK. Some 30 veterinary practices, all of whom use Vet Solutions' software, have "signed up" to the SAVSNET initiative.

The SAVSNET veterinary questionnaires comprise a tabular (tick box) and a free text section. Each questionnaire describes a consultation and is completed by the vet conducting the consultation. In the tabular section of the questionnaire, specific questions regarding certain veterinary conditions are asked (e.g. presence of the condition, severity, occurrence, duration), these questions define the hierarchy of classes. An example (also used for evaluation purposes in Section 5) is presented in Figures 4, 5, 6 and 7[1]. As shown in Figure 4, the first level in the hierarchy (in this example) distinguishes between GI (gastrointestinal) symptoms, namely: "diarrhoea" (D), "vomiting" (V) and "vomiting & diarrhoea" (V&D). The second level (Figure 5) distinguishes between the severity of the GI symptom presented: "haemorrhagic" (H), "non haemorrhagic" (NH) and "unknown severity" (US). The nodes at the next level (Figure 6) consider whether the identified symptom is: "first time" (1st), "nth time" (Nth) or "unknown occurrence" (UO). Finally (Figure 7), the nodes of the fourth level relate to the duration of the symptom: "less than one day" (<1), "between two and four days" (2-4), "between five and seven days" (5-7), "more than eight days" (8+) and "unknown duration" (UD). In this example, the child nodes for each of the GI symptoms are similar, making them hierarchically symmetric. However, our proposed method will work equally well on asymmetric hierarchies.

The tabular section of the questionnaires also includes attributes that are associated with general details concerning the consultation (e.g. date, consultation ID, practice ID), while others are concerned with the "patient" (e.g. species, breed, sex) and its owner (e.g. postcode). The classification/summarisation of the tabular element of the SAVSNET questionnaires is not the topic of interest with respect to this paper; this paper is concerned

[1] It should be noted that levels in the hierarchy are presented in separate figures for convenience only, they are in fact connected and should not be viewed as being independent.

with the summarisation of the free text element of the questionnaires. The free text section of the questionnaires usually comprises notes made by vets, which typically describe the symptoms presented, the possible diagnosis and the treatment to be prescribed. It is the free text section that we are interested in summarising, although in some cases the free text element of the questionnaires is left blank.

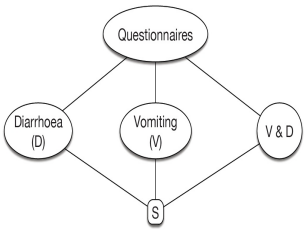

Fig. 4: Level 1, GI symptoms.

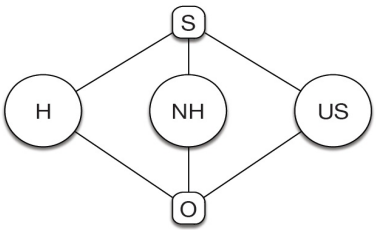

Fig. 5: Level 2, "Severity".

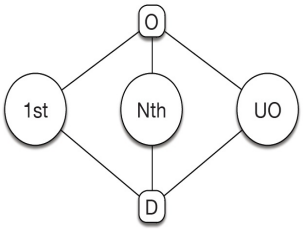

Fig. 6: Level 3, "Occurrence".

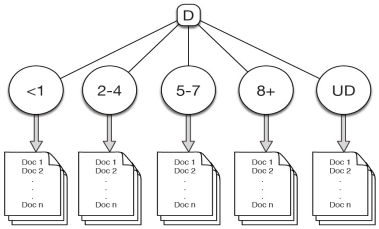

Fig. 7: Level 4, "Duration".

5 Classifier Generation and Application

In this section the two hierarchical text classifier generation approaches considered (cascading and non-cascading) are described in more detail. Recall that in the non-cascading approach the classification process is carried out independently in each node and, as its name implies, independently of the levels and the parent-child node relationship, in other words, flat classifiers are generated for each node; in the cascading approach the output of the classification of the parent nodes affects the classification of child nodes at the next level of the hierarchy. In both cases a 6 step classifier generation process is specified, as follows:

1. **Preprocessing of documents:** text is converted to lower case; numbers, symbols and stop words (common words that are not significant for the text classification/summarisation process) are removed, stemming is applied using an implementation of the Porter Stemming algorithm [31] and feature selection is performed using an implementation of the chi-square method [32].
2. **Classification of documents:** a classifier is generated for each node in the current level of the hierarchy. The classifier generation method chosen with respect to the evaluation included in this paper is the Support Vector Machine (SVM) method, via an implementation of Platt's

Sequential Minimal Optimization (SMO) algorithm [20]. The nature of the classification depends on the approach taken:

(i) Cascading approach: The output of the classification of the nodes in the current level will affect nodes at the next level down in the hierarchy.

(ii) Non-cascading approach: The classifier generation is carried out independently in each node. The classification results do not affect the classifier generation in nodes at the next level down in the hierarchy.

3. **Evaluation of the classification:** The generated classifier is evaluated and the results recorded. Based on the resulting evaluation metrics:

   (i) Cascading approach: Correctly and incorrectly classified instances are considered for the classification process in the next level down in the hierarchy.

   (ii) Non-cascading approach: The classification results from the previous level are not taken into account for the classifier generation conducted at the lower level down in the hierarchy.
4. **Verification of the existence of nodes in the next level down in the hierarchy:** Exit (hierarchical classifier is complete), if there are no more nodes to be developed at the next level down in the hierarchy. Otherwise continue with step 5.
5. **Classifier generation at the next level down in the hierarchy:** repeat process from step 2 for each node at the next level down in the hierarchy.

   (i) Cascading approach: Correctly and incorrectly classified instances for nodes in the previous level are taken into account for nodes in the current level.

   (ii) Non-cascading approach: Classifier generation at nodes in the current level will be performed regardless of the classification results produced at the previous level.

Once complete (and found to be effective when applied to an appropriately defined test set) the generated classifier may be applied to unseen data. A summary for each document is then produced using the resultant class labels generated at each level of the hierarchy. An example of such a summarisation, with respect to the text application considered in the following section, might be: {*Presented diarrhoea, not haemorrhagic, presented for the first time and the duration of the symptoms was between two and four days.*} Note that such a summarisation differs from traditional text summarisation techniques in that the words or phrases that comprise the resulting summary are not necessarily present in the original text. However the resulting summary is a concise, coherent and informative overview of the content of a document.

6 Evaluation

The evaluation of the proposed hierarchical technique for generating summaries from free text was carried out using a subset of the SAVSNET questionnaire corpus which we called SAVSNET-917; for the evaluation we also concentrated on summarising symptoms. This dataset is comprised of 917 records and several classes arranged over four different levels that were defined by specific questions, included in the questionnaire collection process, regarding certain veterinary conditions. The hierarchical arrangement of class labels is as shown in Figures 4, 5, 6 and 7 (described previously). Despite having a relatively small number of records in the SAVSNET-917 dataset, the four levels of the hierarchy were adequately taken into account for the experiments. The distribution of the documents per class for the four levels of the hierarchy is shown in Tables 1, 2, 3 and 4. From these tables it can be seen that the distribution of the documents per class is significantly unbalanced.

Table 1: Number of records per class in the first level of SAVSNET-917 hierarchy.

Class	Num.
Diarrhoea	536
Vomiting	248
Vom&Dia	133
Total	917

Table 2: Number of records per class in the second level of SAVSNET-917 hierarchy.

Class	Num.
Haemorrhagic	177
NotHaemorrhagic	604
UnknownSeverity	136
Total	917

Table 3: Number of records per class in the third level of SAVSNET-917 hierarchy.

Class	Num.
FirstTime	573
NthTime	290
UnknownOccurrence	54
Total	917

Table 4: Number of records per class in the fourth level of SAVSNET-917 hierarchy.

Class	Num.
LessThanOneday	273
BetweenTwoAndFourDays	411
BetweenFiveAndSevenDays	82
MoreThanEightDays	139
UnknownDuration	12
Total	917

The evaluation was conducted using Ten-fold Cross Validation (TCV). The evaluation metrics used were overall accuracy (Acc) expressed as a percentage, Area Under the receiver operating Curve (AUC) [13], sensitivity (Sn) and specificity (Sp). In relation to a confusion matrix, sensitivity measures the proportion of actual positives which are correctly identified, and specificity measures the proportion of negatives which are correctly identified. The AUC measure was used because it takes into consideration the "class priors" (the potential imbalanced nature of the input datasets).

Comparison with other types of summarisation tools, such as a NLP summariser, was not undertaken because of the nature of the different summaries produced; it did not make sense to compare a summary produced in the form of (say) a collection of keywords with a summary produced using our proposed hierarchical classification approach. We could have compared the operation of our approach with the result produced by the application of a sequence of "flat" classifiers, but this would simply have mimicked the

operation of our non-cascading approach; thus such comparisons are not reported here. What we can say is that the validity of our summaries has been confirmed by domain experts working on the SAVSNET project.

Table 5 shows the results for the first level, which were the same regardless of the approach used because there was no parent level that had an influence on the classification process. The results for the other levels are shown in Tables 6, 7 and 8. The cascading and the non-cascading approaches are indicated in the tables using the abbreviations casc and ¬casc respectively.

Table 5: Classification results for level 1.

	Level 1			
Approach	Acc (%)	AUC	Sn	Sp
Both approaches	71.32	0.743	0.713	0.734

Table 6: Classification results for level 2.

Level 2	Diarrhoea				Vomiting				Vomiting and diarrhoea			
	Acc (%)	AUC	Sn	Sp	Acc (%)	AUC	Sn	Sp	Acc (%)	AUC	Sn	Sp
casc	60.81	0.60	0.61	0.54	75.57	0.50	0.76	0.23	59.26	0.62	0.59	0.65
¬casc	73.69	0.64	0.74	0.55	89.11	0.49	0.89	0.10	98.50	0.50	0.98	0.02

Table 7: Classification results for level 3.

Level 3	Haemorrhagic				Not Haemorrhagic				Unknown Severity			
	Acc (%)	AUC	Sn	Sp	Acc (%)	AUC	Sn	Sp	Acc (%)	AUC	Sn	Sp
casc	62.73	0.59	0.63	0.54	62.58	0.61	0.63	0.58	67.19	0.58	0.67	0.49
¬casc	62.71	0.60	0.63	0.58	65.23	0.60	0.65	0.54	58.82	0.60	0.59	0.62

Table 8: Classification results for level 4.

Level 4	First Time				Nth Time				Unknown Occurrence			
	Acc (%)	AUC	Sn	Sp	Acc (%)	AUC	Sn	Sp	Acc (%)	AUC	Sn	Sp
casc	50.08	0.60	0.50	0.66	44.28	0.63	0.44	0.72	–	–	–	–
¬casc	49.21	0.59	0.49	0.66	45.86	0.65	0.46	0.75	51.85	0.44	0.52	0.35

In both approaches no incorrectly classified instances were removed, so the overall number of documents in each level is the same during the experiments. In the case of the cascading approach the instances that were correctly classified and were considered for the classifier generation in the next level down in the hierarchy improved the quality of the resulting summaries; providing for completeness, accuracy and coherency. However, a drawback of this approach is that incorrectly classified documents from a parent node will affect the resulting classification at child nodes and therefore the quality of the summaries produced. In the case of the non-cascading approach the generation of a classifier in each node is in isolation and only the number, quality and distribution of the instances per class will affect the quality of the classifier.

As can be seen from the results shown in Tables 5, 6, 7, and 8, while the accuracy increased from the first to the second level of the hierarchy, it decreased in the third and fourth levels. It can be conjectured that the unbalanced distribution of training texts per class effected the performance for both approaches and in the case of the cascading method the propagation of errors from parent to child was also found to have a considerable impact in the performance of the hierarchical text classification. If incorrectly classified records had been removed the accuracy and AUC values would have been increased from the higher to the lower nodes in the hierarchy due to not having wrongly classified records to consider. However it was conjectured that the removal of wrongly classified records might result in overfitting. Incorrectly classified documents were therefore not removed.

7 Conclusions and Future Work

This paper has presented an approach to the generation of text summaries using a hierarchical text classification approach. The main advantages of the proposed approach, over other text summarisation techniques, are: (i) a more intuitive way of understanding the contents of a document, because humans are used to the concept of defining things in a hierarchical way; and (ii) the ability to handle large document sets due to the hierarchical approach's inherent support for the divide-and-conquer strategy.

The approach was tested using the free text element of a subset of the SAVSNET dataset (SAVSNET-917). The reported experiments were carried out considering a four level hierarchy. The hierarchical classification was performed using two approaches: cascading and non-cascading. In the former approach the performance of a classifier at a parent node influenced the performance at its child nodes, in the latter case each node was considered independent of each other.

For evaluation purposes a Support Vector Machine (SVM) classifier implementation using Platt's Sequential Minimal Optimization (SMO) algorithm was adopted, but other types of classifier generator could equally well be used. The reported evaluation was conducted using TCV. We used a number of evaluation metrics so as to present a wide oversight of the performance of the proposed approaches: accuracy, AUC, sensitivity and specificity were used. The technique was evaluated in terms of the performance of the text classification because the summaries are generated with the labels found in the nodes of each level of the hierarchy. In other words, the generated summaries depend on how well the hierarchical text classification process performs. Besides the evaluation results of the proposed approach, domain experts reviewed the completeness, content accuracy (how accurate is the summary with respect of the original text) and coherency of the summaries generated. Although the domain experts reported that the summaries were complete and coherent, the accuracy of their content is expected to improve with a better performance of the hierarchical classification strategies. Results showed that both the cascading and the non-cascading hierarchical classifi-

cation approaches performed relatively well when a considerable number of records were held at a given node. However, there is still work to be done with respect to the situation where we have few records at a node.

For future work, we also intend to consider extending the proposed technique to address multi-label classification at each hierarchy level in order to produce more comprehensive summaries than using just one label per hierarchy level. It may also be of interest to include the tabular component of the questionnaires in the hierarchical classification process so as to extend and improve the technique. Future work will also consider the application of the proposed technique to several other data sets (including benchmark data sets); a subsequent comparison of the obtained results will be carried out in order to evidence the generality of the technique. Extensive experiments comparing the proposed technique to other text summarisation techniques are also planned, although this will require derivation of appropriate comparison metrics.

References

1. Afantenos, S. and Karkaletsis, V. and Stamatopoulos, P. (2005). Summarization from medical documents: a survey. Artificial Intelligence in Medicine Vol. 33, pp157-177.
2. Alonso, L. and Castellón, I. and Climent, S. and Fuentes, M. and Padró, L. and Rodríguez, H. (2004). Approaches to text summarization: Questions and answers. Inteligencia Artificial Vol. 8, pp22.
3. Celikyilmaz, A. and Hakkani-Tür, D. (2011). Concept-based classification for multi-document summarization. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011, pp5540-5543.
4. Chuang, W. and Tiyyagura, A. and Yang, J. and Giuffrida, G. (2000). A fast algorithm for hierarchical text classification. Data Warehousing and Knowledge Discovery, pp409-418.
5. Dhillon, I.S. and Mallela, S. and Kumar, R. (2002). Enhanced word clustering for hierarchical text classification. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp191-200.
6. Dumais, S. and Chen, H. (2000). Hierarchical classification of web content. Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pp256-263.
7. Duwairi, R. and Al-Zubaidi, R. (2011). A Hierarchical K-NN Classifier for Textual Data. The International Arab Journal of Information Technology. Vol. 8, pp251-259.
8. Fragoudis, D. and Meretakis, D. and Likothanassis, S. (2005). Best terms: an efficient feature-selection algorithm for text categorization. Knowledge and Information Systems. Vol. 8, pp16-33.
9. Gao, F. and Fu, W. and Zhong, Y. and Zhao, D. (2004). Large-Scale Hierarchical Text Classification Based on Path Semantic Vector and Prior Information. CIS'09. International Conference on Computational Intelligence and Security. Vol. 1, pp54-58.
10. Garcia-Constantino, M. F. and Coenen, F. and Noble, P. and Radford, A. and Setzkorn, C. and Tierney, A. (2011). An Investigation Concerning the Generation of Text Summarisation Classifiers using Secondary Data. Seventh International Conference on Machine Learning and Data Mining. Springer, pp387-398.
11. Garcia-Constantino, M. F. and Coenen, F. and Noble, P. and Radford, A. and Setzkorn, C. (2012). A Semi-Automated Approach to Building Text Summarisation Classifiers. To be presented at the Eight International Conference on Machine Learning and Data Mining. Springer.

12. Granitzer, M. (2003). Hierarchical text classification using methods from machine learning. Master's Thesis, Graz University of Technology.
13. Hand, D.J. and Till, R.J. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Machine Learning, 45, pp171-186.
14. Hardy, H. and Shimizu, N. and Strzalkowski, T. and Ting, L. and Zhang, X. and Wise, G.B. (2002). Cross-document summarization by concept classification. Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp121-128.
15. Jaoua, M. and Hamadou, A. (2003). Automatic text summarization of scientific articles based on classification of extracts population. Computational Linguistics and Intelligent Text Processing, pp363-377.
16. Jones, K.S. and others. (1999). Automatic summarizing: factors and directions. Advances in automatic text summarization, pp1-12.
17. Katakis, I. and Tsoumakas, G. and Vlahavas, I. (2008). Multilabel text classification for automated tag suggestion. Proceedings of the ECML/PKDD 2008. Workshop in Discovery Challenge, pp75-83. Antwerp, Belgium.
18. Koller, D. and Sahami, M. (1997). Hierarchically Classifying Documents Using Very Few Words. Proceedings of the Fourteenth International Conference on Machine Learning, pp170-178.
19. Kumilachew, A. (2011). Hierarchical Amharic News Text Classification: Using Support Vector Machine Approach. VDM Verlag Dr. Müller.
20. Platt, J.C. (1999). Using analytic QP and sparseness to speed training of support vector machines. Advances in neural information processing systems, pp557-563.
21. Pulijala, A. and Gauch, S. (2004). Hierarchical text classification. International Conference on Cybernetics and Information Technologies, Systems and Applications: CITSA, pp21-25.
22. Qiu, X. and Huang, X. and Liu, Z. and Zhou, J. (2011). Hierarchical Text Classification with Latent Concepts. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Vol. 2, pp598-602.
23. Radford, A. and Tierney, Á. and Coyne, K.P. and Gaskell, R.M. and Noble, P.J. and Dawson, S. and Setzkorn, C. and Jones, P.H. and Buchan, I.E. and Newton, J.R. and Bryan, J.G.E. (2010). Developing a network for small animal disease surveillance. Veterinary Record. Vol. 167, pp472-474.
24. Rousu, J. and Saunders, C. and Szedmak, S. and Shawe-Taylor, J. (2005). Learning Hierarchical Multi-Category Text Classification Models. Proceedings of the 22nd International Conference on Machine Learning, pp744-751.
25. Ruiz, M.E. and Srinivasan, P. (2002). Hierarchical text categorization using neural networks. Information Retrieval. Vol. 5, pp87-118.
26. Saravanan, M. and Raj, P.C.R. and Raman, S. (2003). Summarization and categorization of text data in high-level data cleaning for information retrieval. Applied Artificial Intelligence, Vol. 17, pp461-474.
27. Sebastiani, F. (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR). Vol. 34, pp1-47.
28. Silla, C.N. and Freitas, A.A. (2011). A survey of hierarchical classification across different application domains. Data Mining and Knowledge Discovery Vol. 22, pp31-72.
29. Sun, A. and Lim, E.P. (2001). Hierarchical text classification and evaluation. ICDM 2001, Proceedings IEEE International Conference on Data Mining. IEEE, pp521-528.
30. Toutanova, K. and Chen, F. and Popat, K. and Hofmann, T. (2001). Text classification in a hierarchical mixture model for small training sets. Proceedings of the tenth international conference on Information and knowledge management, pp105-113.
31. Willett, P. (2006). The Porter stemming algorithm: then and now. Program: electronic library and information systems Vol. 40, pp219-223.
32. Zheng, Z. and Wu, X. and Srihari, R. (2004). Feature selection for text categorization on imbalanced data. ACM SIGKDD Explorations Newsletter Vol. 6, pp80-89.