# An Ensemble Model for Automatic Grading of Evidence

**Yuting Guo** and **Yao Ge**
Computer Science
Emory University
Atlanta GA 30322, USA
yuting.guo@emory.edu
yao.ge@emory.edu

**Ruqi Liao**
Industrial and System Engineering
Georgia Institute of Technology
Atlanta GA 30332, USA
rliao34@gatech.edu

**Abeed Sarker**
Biomedical Informatics
Emory University
Atlanta GA 30322, USA
abeed@dbmi.emory.edu

## Abstract

This paper describes our approach for the automatic grading of evidence task from the Australasian Language Technology Association (ALTA) Shared Task 2021. We developed two classification models with SVM and RoBERTa and applied an ensemble technique to combine the grades from different classifiers. Our results showed that the SVM model achieved comparable results to the RoBERTa model, and the ensemble system outperformed the individual models on this task. Our system achieved the first place among five teams and obtained 3.3% higher accuracy than the second place.

## 1 Introduction

Evidence-based medicine (EBM) requires the making of clinical decisions using the current best external evidence rather than solely relying on clinical experience and pathophysiologic rationale (Sackett et al., 1996). To adhere to EBM best practice, practitioners need to identify the best quality evidence associated with a clinical query. To grade the quality of evidence, Ebell et al. (2004) proposed the Strength of Recommendation Taxonomy (SORT). SORT has a three-levels for rating—A (strong), B (moderate), and C (weak), where A-level is based on high-quality studies with consistent results; B-level is based on high-quality studies with inconsistent results or some limitations; C-level is based on the studies with severe limitations. It is a straightforward grading system that allows clinical experts to rate individual studies or bodies of evidence based on quantity, quality, and consistency.

To address the challenging problem of automatically grading the quality of evidence, the Australasian Language Technology Association (ALTA) Shared Task 2021 organized a competition. The participants were required to develop a system to predict the grade of evidence given multiple related medical publications. Our team trained several supervised classifiers to address the problem. Our approach included traditional supervised classification models such as support vector machines (SVM) (Cortes and Vapnik, 1995), neural network models using pretrained models (RoBERTa) (Liu et al., 2019), and an innovative ensemble system which combines the predictions of multiple classifiers. Our results showed that the SVM model achieved comparable results to the RoBERTa model, and the ensemble system outperformed the individual models on this task. The ensemble model combines the prediction from multiple classifiers in a unique manner: grades (A, B or C) predicted by each classifier is first converted into a continuous number, and then all the numbers are added for each instance. Using the training data, the best separations for the numeric totals are computed. These numeric boundaries are then used to convert continuous scores in the test set to discrete evidence grades. Our system achieved the first place among five teams and obtained 3.3% higher accuracy than the second place.

## 2 Related Work

ALTA Shared Task 2021 is a re-visit of ALTA Shared Task 2011 (Molla and Sarker, 2011). Previous studies have developed several SVM-based systems for this task. Molla and Sarker (2011) used a sequential approach to combine multiple individual SVM models trained with the features from the titles, body of the abstracts, and publication types. Gyawali et al. (2012) expanded the feature set proposed by Molla and Sarker (2011) with the Medical Subject Headings (MeSH) terms and developed a stacking-based approach to integrate predictions from multiple SVM models. Byczyńska et al. (2020) experimented with a larger set of features and applied multiple machine learning techniques such as classical machine learning mod-

els, neural networks, game theory, and consensus methods. In our work, we trained SVM models on a feature set similar to Byczyńska et al. (2020). We also applied a pre-trained transformer-based model named RoBERTa (Liu et al., 2019), which has achieved state-of-the-art results in a wide range of natural language processing (NLP) tasks.

## 3 Data Description

The data for this shared task consisted of a set of evidence grades under the SORT criteria and a list of related publications associated with each evidence grade. The publications were obtained from PubMed [1] and were provided in the form of XML files which contained the title, the abstract, and some meta-data (*eg.*, publication types, MeSH terms). Some data statistics are shown in Table 1.

| | Train (%) | Dev (%) | Test (%) |
|---|---|---|---|
| A | 31.3 | 27.0 | 30.6 |
| B | 45.9 | 44.9 | 48.6 |
| C | 22.7 | 28.1 | 20.8 |
| Total size | 677 | 178 | 183 |

Table 1: The distribution of the three grades and data set sizes for the training, development, and test sets.

## 4 Method

### 4.1 SVM

We implemented the SVM models with Python 3.7 and the sklearn tool (Pedregosa et al., 2011). We trained multiple SVM models using different feature sets for each, which included the number of related publications (*npmid*), journal titles, and other features, as follows:

**N-gram Features (*n-gram*)** The n-gram features were generated from the texts of the titles and the bodies of the abstracts. Because one evidence grade can be based on multiple publications, we combined the titles and the abstracts of all publications to create sequences of titles and abstracts per evidence, respectively. Then, we computed the term frequency–inverse document frequency (TF-IDF) features from the n-grams ($n = 1, 2, 3, 4$) of the combined sequences.

**Consistency Features (*cons*)** As mentioned in Ebell et al. (2004), the consistency of experimental results can affect the evidence strength. Inspired by that, we detected the mentions of consistent results in the body of abstracts by keyword matching. For each evidence, if any of the publications matched the word "consistent" or "consistency" in the abstract, the consistency feature was set as 1; otherwise it was set to 0.

**Publication Types (*pubtype*)** As discussed in Molla and Sarker (2011) and Byczyńska et al. (2020), publication types can be a strong indicator of the evidence strength. We extracted the publication type terms tagged as *PublicationType* in the XML files and assigned a pseudo publication type "unknown" to the publications without any *PublicationType* tag. In addition, we used the PubMed tool [2] to retrieve the publication type IDs. We used one-hot encoding to encode the publication type terms and IDs, respectively. Also, we generated a publication type rank according to the level of evidence pyramid in Sarker and Mollá-Aliod (2010). The rank ranged from 0 to 5, where higher number indicates higher quality.

**MeSH** MeSH terms provide information regarding the topics covered in a publication. We used the PubMed tool to request MeSH term IDs and represented the MeSH feature by one-hot encoding.

### 4.2 RoBERTa

Encouraged by the success of the pre-trained transformer-based models in recent years, we developed a classifier using RoBERTa, one of the most popular pre-trained transformer-based models. The classification model architecture was the same as the model in (Liu et al., 2019). It consists of an encoder, which converted the input text sequence into an embedding vector, and a classification layer with softmax activation, which projected the embedding vector into a class probability vector. The inputs were the abstract texts of the publications associated with each evidence instance. However, if we attached the abstracts into one sequence, the input length often exceeded the maximum sequence length limitation of RoBERTa, which is 512 characters. Therefore, we re-organized the dataset by splitting the evidences involving multiple publications into different instances so that each instance only contained one evidence and one publication, as shown in Figure 1. During the inference phase,

[1] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[2] https://www.ncbi.nlm.nih.gov/pmc/tools/get-metadata/

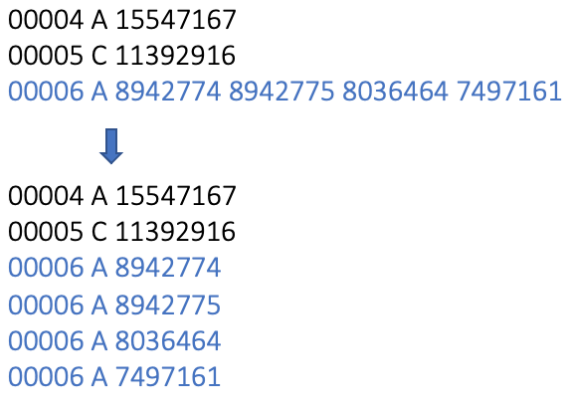


Figure 1: An example of the data re-organization process. The first column contains the evidence IDs, the second column contains the SORT grades, and the third column contains the publication IDs.

for each evidence, the class probability vectors of multiple publications were averaged, and the class with the highest probability was chosen as the final prediction.

### 4.3 Ensemble

Because the classes A, B and C represent the strength of evidence from strong to weak, we considered the task of grading as a regression problem (rather than a classification problem) and converted the predictions from the classifiers into numbers on a numeric scale. Specifically, we represented the classes A, B, and C as the numbers 0, 1, and 2. For each instance, we computed a numeric score (rather than a discrete category) by adding up the converted predictions from all classifiers. Following this process, we performed grid search to find two thresholds in which the evidences with scores smaller the lower threshold were classified as A, those larger than the higher threshold were classified as C, and those with scores between the lower and upper thresholds were classified as B. Optimal values for the thresholds were based on the training set. In addition, considering the fact that the classifiers with low accuracies may hurt the performance of the ensemble model, we greedily removed the least accurate classifiers to find the classifier set that achieved the best performance on the training/development set.

## 5 Experiments

**SVM** We trained the SVM models for all possible combinations of the features and experimented with not using class weights and using the empirical class weights $W_A = 1.2$, $W_B = 1.2$, and $W_C = 1.0$. In total, we created 127 feature combinations and obtained 254 classification models. For each model, we performed grid search on the development set to find the best configuration for the regularization parameter $C \in \{1, 2, 4, 6, 8\}$ and the kernel type $K \in \{"linear", "rbf"\}$.

**RoBERTa** The specific version of RoBERTa we used was RoBERTa-large. According to the preliminary experiments, we set the batch size as 32, the learning rate as $8 \times 10^{-6}$, and the maximum sequence length as 256. The model was trained for 10 epochs with 3 random initialisations.

For both SVM and RoBERTa, we tuned the parameters based on the training set and the development set to find the optimal parameters, and we re-trained the model with the optimal parameters on the whole data set (i.e., the combination of the training set and the development set). The reported results of the test set were predicted by the models trained on the whole data set, and those of the development set were predicted by the models trained on the training set.

## 6 Results

Table 2 shows the results of the best individual SVM model, the RoBERTa model, and the ensemble model on the development set and the test set. For the SVM model, the best feature combination is *n-gram+pubtype+npmid*. The results show that the performance of the RoBERTa model is comparable to the SVM model, and the ensemble model outperformed the other two models. However, the differences between the three models were not statistically significant according to the 95% confidence intervals. Also, we observed that the performances were considerably lower for the test set compared to the development set. This suggests that the models may overfit on the training/development data because of the small data size.

For further error analysis, we plotted the confusion matrix for our best system (*ie*., the ensemble model), shown in Figure 2. As we can see, the majority of errors can be attributed to the misclassification of the classes A and C. Most A-level and C-level evidences were predicted as B. This can be another indicator of overfitting because the majority evidences in the training set were graded as B.

| Model | Dev | 95% CI | Test | 95% CI |
|---|---|---|---|---|
| SVM | 0.63 | 0.48-0.76 | 0.48 | 0.34-0.60 |
| RoBERTa | 0.58 | 0.44-0.70 | 0.48 | 0.34-0.62 |
| Ensemble | **0.7** | 0.58-0.82 | **0.54** | 0.38-0.68 |

Table 2: The accuracies and 95% confidence intervals (CIs) on the development and test set.

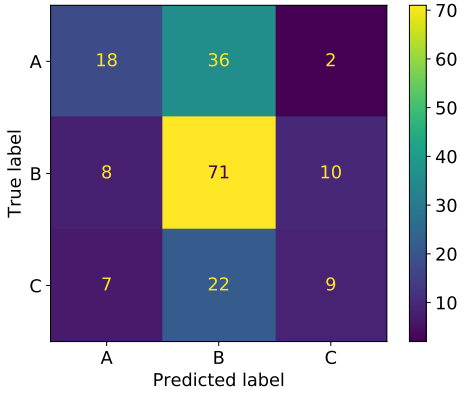


Figure 2: The confusion matrix for the result of the ensemble model on the test set.

## 7 Discussion

As illustrated in Table 6, the RoBERTa model did not outperform the SVM model on this task. This finding is somewhat surprising because many recent studies have shown that pre-trained transformer-based models can achieve the state-of-the-art performance on a wide range of natural language processing tasks (Liu et al., 2019; Devlin et al., 2019; Nguyen et al., 2020; Yang et al., 2019). A possible explanation for this can be that the most important factor for the evidence strength grading is the publication type and the consistency of the experiments (Ebell et al., 2004). In our experiments, the input for RoBERTa was only the abstracts, which rarely contained the publication type information. In contrast, in the abstracts, the consistency of the experiments are usually implicitly described by comparing the experimental results which involve numbers. It has been suggested that the pre-trained transformer-based models lack in the ability of effectively representing numbers (Wallace et al., 2019). Therefore, further studies will need to be undertaken to explore how to incorporate the meta-data information into transformer-based models and how to make such models understand/compare numbers.

Although we achieved the top place in this competition, some systems described in past publications achieved higher accuracies than our best result (Molla and Sarker, 2011; Gyawali et al., 2012; Byczyńska et al., 2020). We noted that all of these systems used the publication type features. Moreover, Byczyńska et al. (2020) showed that using the single publication type feature achieved 70% accuracy on the test set. However, in our experiments, our model with the single publication type feature only achieved 52% accuracy. We speculate that the cause of the performance gap might be due to the fact that we processed the publication type feature differently compared to the abovementioned publication. In our method, we simply used the publication type terms extracted from the XML files, while Byczyńska et al. (2020) used a rule-based system to identity the publication types from the titles and the abstracts. Further research is needed to explore effective methods for processing the publication type feature.

## References


Aleksandra Byczyńska, Maria Ganzha, Marcin Paprzycki, and Mikołaj Kutka. 2020. Evidence quality estimation using selected machine learning approaches. In *2020 Conference on Information Communications Technology and Society (ICTAS)*, pages 1–8.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

M. H. Ebell, J. Siwek, B. D. Weiss, S. H. Woolf, J. Susman, B. Ewigman, and M. Bowman. 2004. Strength of recommendation taxonomy (SORT): a patient-centered approach to grading evidence in the medical literature. *J Am Board Fam Pract*, 17(1):59–67.

Binod Gyawali, Thamar Solorio, and Yassine Benajiba. 2012. Grading the quality of medical evidence. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 176–184, Montréal, Canada. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv*, 1907(11692).

Diego Molla and Abeed Sarker. 2011. Automatic grading of evidence: the 2011 ALTA shared task. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 4–8, Canberra, Australia.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A Pre-trained Language Model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

David L Sackett, William M C Rosenberg, J A Muir Gray, R Brian Haynes, and W Scott Richardson. 1996. Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71–72.

Abeed Sarker and Diego Mollá-Aliod. 2010. A rule based approach for automatic identification of publication types of medical papers. In *Proceedings of the ADCS Annual Symposium*, pages 84–88. Citeseer.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP models know numbers? probing numeracy in embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5307–5315, Hong Kong, China. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc.