Overview of the 2022 ALTA Shared task: PIBOSO sentence classification, 10 years later

Diego Mollá
School of Computing
Macquarie University
Sydney, Australia
diego.molla-aliod@mq.edu.au

Abstract

The 2022 ALTA shared task has been running annually since 2010. This year, the shared task is a re-visit of the 2012 ALTA shared task. The purpose of this task is to classify sentences of medical publications using the PIBOSO taxonomy. This is a multi-label classification task which can help medical researchers and practitioners conduct Evidence Based Medicine (EBM). In this paper we present the task, the evaluation criteria, and the results of the systems participating in the shared task.

1 Introduction

Within the practice of Evidence Based Medicine (EBM), the medical practitioner integrates individual clinical expertise with the best external evidence at point of care (Sackett et al., 1996). Finding the best available evidence, however, is increasingly difficult given the large amount of medical publications. For example, at the time of writing, PubMed contains more than 34 million citations for biomedical literature[1]. From 2020 to present, CORD-19, a resource of medical publications about COVID-19, SARS-COV-2, and related coronaviruses, has increased from an initial set of 28,000 papers (Wang et al., 2020) to over 1,000,000[2].

To assist with the task of finding the best available evidence, best EBM practice suggests users to formulate queries that focus on specific aspects of the clinical information sought. PIBOSO (Kim et al., 2011) is a pre-defined set of such aspects of clinical information, and systems participating in the 2012 ALTA shared task classified sentences from medical publications into PIBOSO labels (Amini et al., 2012). In 2022, 10 years later, ALTA has re-visited the task, to find out whether recent advances in machine learning would allow to improve the quality of such classifiers.

This paper presents the results of systems participating in the 2022 ALTA shared task. Section 2 describes the PIBOSO taxonomy. Section 3 briefly mentions related work between the 2012 and the 2022 ALTA shared tasks. Section 4 describes the evaluation framework. Section 5 presents two simple baselines that were made available to the participating teams. Section 6 presents the results and briefly describes the methods of participating systems. Finally, Section 7 concludes this paper.

2 PIBOSO

EBM guidelines recommend the use of structured queries that focus on specific aspects of clinical information (Richardson et al., 1995). One of the most widely used systems is PICO, which defines 4 types of information: ***P****opulation*, for example the number and type of participants in a study; ***I****ntervention*, such as the treatment applied to the population; ***C****omparison* (if appropriate), for example alternative interventions or placebo; and ***O****utcome* of an intervention.

Different variants and extensions of PICO have been proposed. The ALTA 2012 and 2022 shared tasks use PIBOSO (Kim et al., 2011). This schema removes the ***C****omparison* tag and adds three new tags: ***B****ackground*, ***S****tudy design*, and ***O****ther*. The PIBOSO tags, as defined by Kim et al. (2011), are:

- ***P****opulation*: The group of individual persons, objects, or items comprising the study's sample, or from which the sample was taken for statistical measurement;
- ***I****ntervention*: The act of interfering with a

[1] https://pubmed.ncbi.nlm.nih.gov/, accessed on 15 November 2022.

[2] https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge, accessed on 15 November 2022.

condition to modify it or with a process to change its course (includes prevention);

- ***B**ackground*: Material that informs and may place the current study in perspective, e.g. work that preceded the current; information about disease prevalence; etc;
- ***O**utcome*: The sentence(s) that best summarizes the consequences of an intervention;
- ***S**tudy Design*: The type of study that is described in the abstract;
- ***O**ther*: Any sentence not falling into one of the other categories and presumed to provide little help with clinical decision making, i.e. non-key or irrelevant sentences.

Different parts of a medical publication may focus on different PIBOSO elements. In practice, each sentence of a PubMed abstract will normally focus on one PIBOSO element, but sometimes a sentence may focus on several (see Table 2 for examples). Thus, systems attempting to determine the PIBOSO labels of a sentence will need to implement multi-label sentence classification. This is the focus of the 2012 and 2022 ALTA shared tasks.

3 Related Work: From 2012 to 2022

The data used in this 2022 shared task is based on the data from the 2012 task (Amini et al., 2012), which is derived from the original NICTA-PIBOSO dataset by Kim et al. (2011). Sentence classification systems participating in ALTA 2012 used approaches based on Conditional Random Field (CRF), Support Vector Machines (SVM), stacked logistic regression, maximum entropy, and random forests. The results of the participating systems are summarised in Table 1.

The following additional research has used the NICTA-PIBOSO dataset for sentence classification. Verbeke et al. (2012) used statistical relational learning. Hassanzadeh et al. (2014) used CRF and a discriminative set of features. Jin and Szolovits (2020) used LSTM plus adversarial training and unsupervised pre-training over large corpora. All of these systems report F1 as the evaluation metric, which is different from the metric used in the ALTA 2012 and ALTA 2022 datasets (Section 4). Even though the F1 and AUC metrics may lead to similar rankings of systems, as observed in the ALTA 2012 shared task (Amini et al., 2012), systems fine-tuned for AUC might not lead to optimal F1 scores. Most notably, systems fine-tuned for AUC do not need to set a classification threshold, and an evaluation using F1 will give very different results depending on the choice of classification threshold.[3]

4 Evaluation Framework

We have been unable to retrieve the labelled test data of the 2012 ALTA shared task. As a consequence, the data for the 2022 shared task is based on the training data from the 2012 shared task, after shuffling the original data and re-numbering the sample IDs. The resulting data has been split into three sets for training, validation, and test.

The documents used in the datasets are abstracts of medical publications published in PubMed. Each abstract contains multiple sentences, and consequently a single PubMed abstract corresponds to several samples in the dataset. To minimise data leakage between the different partitions, the partitions were made based on the abstracts so that all sentences of the same abstract would be in the same partition. Besides preventing data leakage, this partitioning also allows the participating systems to use the context of the other sentences from an abstract during the classification task.

Table 2 shows several samples from the dataset. The table shows that the dataset indicates the PubMed ID, the sentence position in the PubMed abstract, the PIBOSO labels associated with the sentence, and the text of the sentence.

Table 3 shows that the label distributions are not balanced, and most of the labels are ***B**ackground*, ***O**utcome*, or ***O**ther*. All three partitions have a similar label distribution.

The evaluation framework was implemented as a CodaLab competition[4] which consisted of three phases. In the **development phase**, the training and validation data were available but the labels of the validation data were not available. Participant teams were able to make up to 100 submissions to test their systems against the validation data. This phase was not used for the final ranking of the participating systems and ended on the 4th of October 2022. In the **test** phase, the test data

[3] We observed that a system participating in ALTA 2022 obtained very good AUC scores but their F1 score was 0 because the probabilities assigned to each label were lower than the default threshold of 0.5. Probably, a lower threshold would have given a non-zero F1 score for that system.

[4] https://codalab.lisn.upsaclay.fr/competitions/6935

System	AUC (test)	F1
Marco Lui (Lui, 2012)	**0.97**	**0.82**
A_MQ	0.96	0.79
Macquarie Test (Molla, 2012)	0.94	0.77
DPMCNA	0.93	0.71
System_Ict (Gella and Long, 2012)	0.93	0.73
Dalibor	0.92	0.73
Starling	0.87	0.79
Mix	0.84	0.74
Benchmarks (Amini et al., 2012)		
- CRF corrected	0.88	0.80
- Naive	0.70	0.55

Table 1: AUC and F1 for the 2012 test set. The best results per column are given in bold. Refer to Section 4 for an explanation of the AUC metric.

PubMed ID	Sentence	Labels	Text
1031546	1	***P**opulation*, ***I**ntervention*	A 26-year-old subfertile woman ...
1031546	2	***O**utcome*	A pregnancy resulted, which ...
1031546	3	***O**utcome*	It is suggested that this production ...

Table 2: Annotations corresponding to one PubMed abstract from the training set

	train	val	test
***P**opulation*	7.11%	7.84%	7.38%
***I**ntervention*	6.10%	6.31%	6.15%
***B**ackground*	21.63%	27.23%	22.67%
***O**utcome*	38.85%	37.25%	35.32%
***S**tudy design*	2.03%	2.61%	2.46%
***O**ther*	29.50%	24.62%	30.75%

Table 3: Label distributions in the data set. The numbers indicate the percentage of sentences that contain the given label. The sum of percentages in each dataset is higher than 100% because a sentence may have multiple labels.

(without labels) was made available and participant teams were able to make up to 3 submissions. This phase was used for the final ranking. In the subsequent phase of **unofficial submissions**, participant systems are able to make an unlimited number of submissions[5] that will be evaluated on the validation data. This phase remains open and new teams are encouraged to participate and make new submissions.[6]

[5]In practice, there is a limit of 999 unofficial submissions.

[6]Read https://codalab.lisn.upsaclay.fr/competitions/6935 and http://www.alta.asn.au/events/sharedtask2022/ for details of how to participate.

The training data contains 8,216 sentences, the validation data used in the first phase contains 459 sentences, and the test data contains 569 sentences.

Given an input sentence, the output of each participating system must produce, for every PIBOSO label, a number between 0 and 1 that represents the confidence or probability that the label is assigned to the sentence.

The evaluation metric is the micro-average of the Area Under the Receiver Operating Characteristic (ROC) Curve. The ROC curve plots the true positive rate against the false positive rate at various threshold settings for binary classification. We use the micro-average so that labels with more samples are given more importance. The advantage of using this metric instead of metrics such as F1 is that it incorporates the probability scores returned by the system, such that two systems with identical classification predictions but different probability scores will be ranked differently.

5 Baselines

We have provided two simple baselines against which the participating systems can compare. The code for these baselines is publicly available[7]. We

[7]https://github.com/altasharedtasks/baselines2022

System	Category	AUC (test)
Heatwave	Student	**0.9874**
CSECU-DSG	Student	0.9687
Cufe	Open	0.9634
TurkNLP	Student	0.9318
NN baseline		0.9105
NB baseline		0.8769

Table 4: Results of the 2022 ALTA shared task. Metric: Area under the micro-averaged Receiver Operator Characteristics (ROC) curve. Sorted based on AUC (test). The winning team is highlighted in **boldface**.

describe these baselines below.

Naive Bayes (NB). A set of 6 independent Naive Bayes classifiers, one per classification label, has been implemented using scikit-learn. Each sentence is vectorised using tf.idf, and the number of features has been limited to 10,000. Stop words are *not* removed.

Neural Network (NN). A simple Neural Network architecture has been implemented in Keras. The sentences have been vectorised in the same way as with the Naive Bayes baseline. Namely, scikit-learn has been used to obtain the tf.idf of the sentences, and the top 10,000 words have been retained. Stop words are *not* removed. The resulting vectors are fed to a simple neural network consisting of a single dense layer with 6 neurons (one per label), and sigmoid activation. The network does not use dropout. The network has been trained for 70 epochs, batch size 32, and a validation split of 0.2. The choice of number of epochs was determined after examining the loss of the validation split[8].

6 Participating Systems and Results

A total of 3 teams registered in the student category, and 6 teams registered in the open category. Of these, only 5 teams submitted runs in the CodaLab test phase. Table 4 shows the results of the baselines and participating systems for the test phase.

We can observe that all participating teams outperformed the two baselines.

Three of the teams have submitted system descriptions and they are available in the proceedings of the 2022 Australasian Language Technology workshop (ALTA 2022). All three systems incorporated Transformers in their implementations, in particular variants of BERT (Devlin et al., 2018).

Team Heatwave obtained the best results. Their winning system (Fang and Koto, 2022) used an ensemble of BERT-based implementations (BERT, RoBERTa, BioBERT) that classified each sentence with the help of the context of adjacent sentences.

Team CSECU-DSG (Aziz et al., 2022) extended DeBERTa with 5-fold cross-training (creating effectively an Ensemble approach) and multi-sample dropout.

Team TurkNLP (Bölücü and Hepsağ, 2022) extended SciBERT by adding a classification layer that incorporated information from the [CLS] token and the average of SciBERT embeddings.

[8]Note that the validation split used for training is part of the ALTA training set. This is different from the actual ALTA validation set.

System	AUC (dev)	AUC (test)
NN baseline	0.9091	0.9105
NB baseline	0.8718	0.8769

Table 5: Results of the baseline systems on the development and test sets. Metric: Area under the micro-averaged Receiver Operator Characteristics (ROC).

7 Conclusions

Participation in the 2022 ALTA shared task showed the successful use of Transformer approaches for this task of multi-label classification of abstract sentences from medical publications using PIBOSO. All participating systems outperformed the baselines. Furthermore, the top system outperformed the participating systems of ALTA 2012 (Tables 1 and 4). There is a potential caveat in that the test data used in the 2022 ALTA shared task was different from that of the 2012 ALTA shared task because of the non-availability of the labels of the original 2012 test data. Having said that, given that the test set of the original data was created as a random partition, we would not expect a very large difference in the results. Table 5 shows very small differences between the results of the development and test sets of the Naive Bayes and Neural Networks baselines. In addition, the 2012 shared task (Amini et al., 2012) showed a difference of 0.01 or less between the public and private test partitions in most participating systems. The small differences in the results suggest that an evaluation made with the 2022 test data would produce similar results to an evaluation made with the 2012 test data.

References

Iman Amini, David Martinez, and Diego Molla. 2012. Overview of the ALTA 2012 shared task. In *Australasian Language Technology Workshop 2012: ALTA Shared Task*.

Abdul Aziz, Md. Akram Hossain, and Abu Nowshed Chy. 2022. Enhanding DeBERTa transformers model for classifying sentences from biomedical abstracts. In *Proceedings of the 2022 Australasian Language Technology Workshop, ALTA 2022*, Adelaide, Australia.

Necva Bölücü and Pinal Uskaner Hepsağ. 2022. Automatic classification of evidence based medicine using transformers. In *Proceedings of the 2022 Australasian Language Technology Workshop, ALTA 2022*, Adelaide, Australia.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Biaoyan Fang and Fajri Koto. 2022. Context-aware sentence classification in evidence-based medicine. In *Proceedings of the 2022 Australasian Language Technology Workshop, ALTA 2022*, Adelaide, Australia.

Spandana Gella and Duong Thanh Long. 2012. Automatic sentence classifier for evidence based medicine: Shared task system description. In *Australasian Language Technology Workshop 2012: ALTA Shared Task*.

Hamed Hassanzadeh, Tudor Groza, and Jane Hunter. 2014. Identifying scientific artefacts in biomedical literature: The evidence based medicine use case. *Journal of Biomedical Informatics*, 49:159–170.

Di Jin and Peter Szolovits. 2020. Advancing PICO element detection in biomedical text via deep neural networks. *Bioinformatics*, 36(12):3856–3862.

S. Kim, D. Martinez, L. Cavedon, and L. Yencken. 2011. Automatic classification of sentences to support evidence based medicine. *BMC bioinformatics*, 12:S5.

Marco Lui. 2012. Feature stacking for sentence classification in evidence-based medicine. In *Australasian Language Technology Workshop 2012: ALTA Shared Task*.

Diego Molla. 2012. Experiments with clustering-based features for sentence classification in medical publications: Macquarie Test's participation in the ALTA 2012 shared task. In *Australasian Language Technology Workshop 2012: ALTA Shared Task*.

W. Scott Richardson, Mark C. Wilson, Jim Nishikawa, and Robert S.A. Hayward. 1995. The well-built clinical question: a key to evidence-based decisions. *ACP J Club*, 123:A12–A13.

David L. Sackett, William M. Rosenberg, Jamuir 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.

Mathias Verbeke, Vincent Van Asch, Roser Morante, Paolo Frasconi, Walter Daelemans, and Luc De Raedt. 2012. A statistical relational learning approach to identifying evidence based medicine categories. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 579–589, Jeju Island, Korea. Association for Computational Linguistics.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex Wade, Kuansan Wang, Nancy Xin Ru Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. Cord-19: The covid-19 open research dataset.