Table 1： A typical example of an abstract with paragraph headings and the corresponding annotation labels. The PMID of this abstract is 28074672. The heading column shows the structured tag from the original abstract. The sentence column shows the corresponding sentence, and the label column shows the tag we define for our own use.

Heading	Label	Sentence
Objective	A	To assess the feasibility of conducting a randomized controlled trial to determine the effectiveness of a twenty-week power…
Design	M	Pilot randomised controlled trial.
Setting	M	A large-scale twenty-four-hour residential facility in the Netherlands.
Subjects	P	Thirty-seven persons with profound intellectual and multiple disabilities.
Intervention	I	Participants in the intervention group received a power-assisted exercise intervention…
Intervention	I	Participants in the control group received care as usual.
Main Measures	O	Trial feasibility by recruitment process and outcomes completion rates…
Results	R	Thirty-seven participants were recruited (M age = 32.1)…
Results	R	Programme compliance rates ranged from 54.2% to 97.7% with a mean (SD) of 81.5% (13.4).
Results	R	Oxygen saturation significantly increased in the intervention group.
Results	R	Standardised effect sizes on the difference between groups in outcome varied between 0.02 and 0.62.
Conclusions	C	The power-assisted exercise intervention and the trial design were feasible and acceptable to people…
Conclusions	C	This pilot study suggests that the intervention improves oxygen saturation…

Table 2: Summary of the numbers and distributions of abstracts and sentences with P/I/O tags.

Label	Articles	Sentences
P	21198	27696
I	13712	24603
O	20473	32526

Table 3: Distributions of P/I/O sentences and non-P/I/O sentences in the processed data sets.

Label	1	0
P	27696	57129
I	24603	60222
O	32526	52299

Note: 1 represents sentences with the specified label, and 0 represents sentences whose labels do not match the specified label.

Table 4: Distributions of P/I/O sentences and non-P/I/O sentences in the training, test, and verification sets.

Label	Train	Test	Dev

	1	0	1	0	1	0
P	22347	46095	2723	5550	2626	5484
I	19865	48557	2231	5942	2507	5723
O	26230	42212	3219	5054	3077	5033

Note: Three data sets were constructed for each model; for example, the values in the P row and the Train column represent the training set for the P classification model, the values in the P row and the Test column represent the test set for the P classification model, and the values in the P row and the Dev column represent the verification set for the P classification model.

Table 5: Word-frequency statistics for PICO sentences.

Sentence type	Top 10 words
P sentences	Patients, years, women, age, group, study, aged, total, hundred, mean
I sentences	Group, patients, mg, received, placebo, weeks, treatment, control, intervention, daily
O sentences	Outcome, primary, scale, measured, months, pain, outcomes, treatment, secondary, assessed

Table 6: Results of n-gram control tests

n-gram range	P_SVM		I_SVM		O_SVM	
	acc	F1	acc	F1	acc	F1
1	0.914	0.863	0.893	0.802	0.911	0.886
2	0.892	0.818	0.863	0.726	0.874	0.831
3	0.809	0.615	0.796	0.483	0.773	0.614
1-2	0.9243	0.8792	0.8992	0.8144	0.9154	0.8913
1-3	0.9237	0.8788	0.8983	0.8139	0.9148	0.8907
2-3	0.893	0.821	0.861	0.721	0.875	0.832

Note: The P_SVM column represents the SVM binary classification model designed for P sentences, and the n-gram range column specifies the parameter(s) of the n-gram model we considered in the TF-IDF analysis, where an n-gram range of 1 corresponds to unigrams and an n-gram range of 2 corresponds to bigrams. We present the acc and F1 values obtained through 10-fold cross-validation as evaluation indicators for comparison. Since the results for n-gram ranges of 1-2 and 1-3 differ by less than 0.1, these results are presented up to four significant digits after the decimal point to allow them to be distinguished. For the six control groups represented in the table, all of the same conditions were used except for the n-gram range, including the data set and all other model parameters.

Table 7: Results of the TF-IDF and word2vec comparison experiment

	P elements			I elements			O elements		
	P	R	F1	P	R	F1	P	R	F1
TF-IDF	0.925	0.838	0.879	0.842	0.789	0.814	0.886	0.897	0.891
word2vec	0.894	0.796	0.842	0.808	0.731	0.768	0.866	0.855	0.861

Note: The experimental results for the TF-IDF method and the word2vec method compared in this study were obtained using the same soft-margin SVM classification model; the P elements column represents the results for a classification model constructed for P sentences, and the I elements column represents the results for a classification model constructed for I sentences.

Table 8: Comparison of the soft-margin SVM model used in this study with the standard RF and XGBoost models.

	P elements			I elements			O elements		
	P	R	F1	P	R	F1	P	R	F1
SVM	0.925	0.838	0.879	0.842	0.789	0.814	0.886	0.897	0.891
RF	0.899	0.813	0.854	0.838	0.759	0.808	0.881	0.834	0.861
XGBoost	0.860	0.857	0.852	0.828	0.814	0.791	0.857	0.838	0.832

Note: The acc, P, R and F1 values obtained through 10-fold cross-validation are used to evaluate the models.

Table 9: Comparison of model results.

	P elements			I elements			O elements		
	P	R	F1	P	R	F1	P	R	F1
NB	0.902	0.925	0.913	0.786	0.716	0.749	0.836	0.920	0.876
LSTM	0.885	0.828	0.856	0.749	0.815	0.781	0.845	0.832	0.838
SVM	0.925	0.838	0.879	0.842	0.789	0.814	0.886	0.897	0.891

Note: The LSTM model was described in the Di Jin 2018 paper, the NB model was described in the Ke-Chun Huang 2011 paper, and the SVM model is the model studied in the present paper.

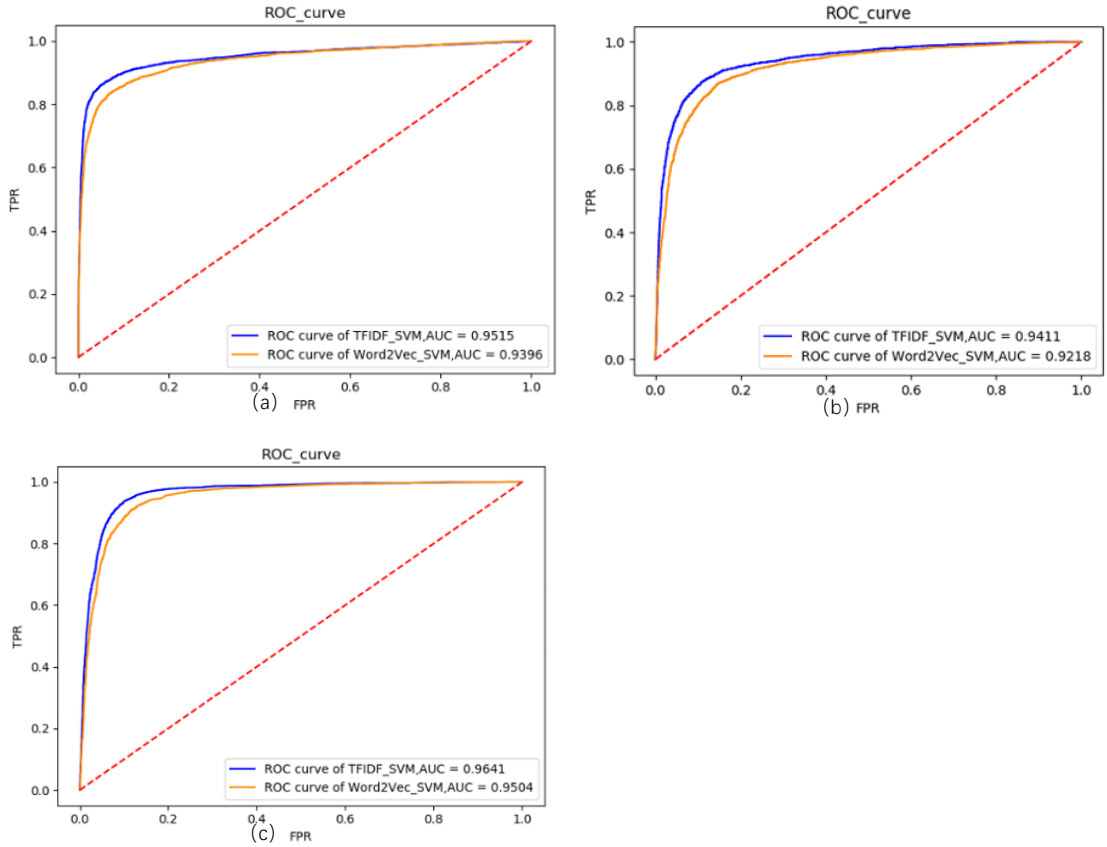

Fig1. The ROC curves of the TF-IDF and word2vec comparison experiment: (a) results for P elements, (b) results for I elements, and (c) results for O elements.

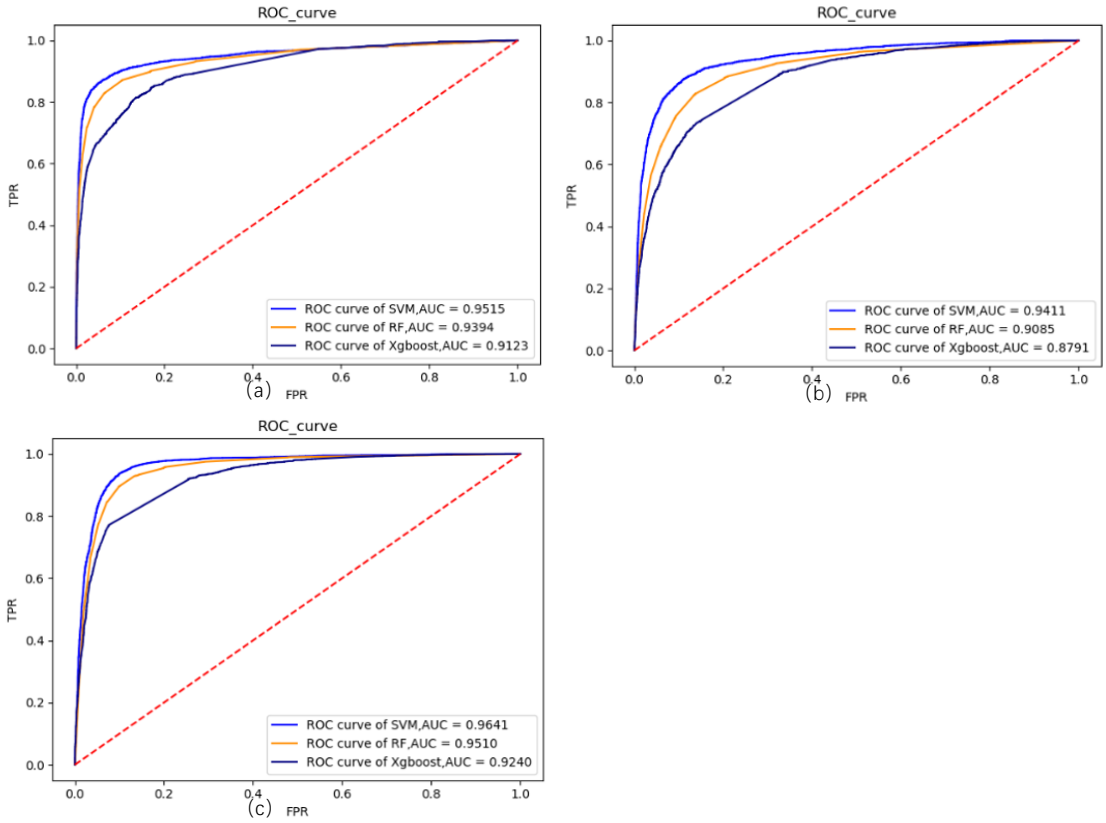

Fig 2: The ROC curves of the SVM model, RF and Xgboost comparison experiment: (a) results for P elements, (b) results for I elements, and (c) results for O elements.

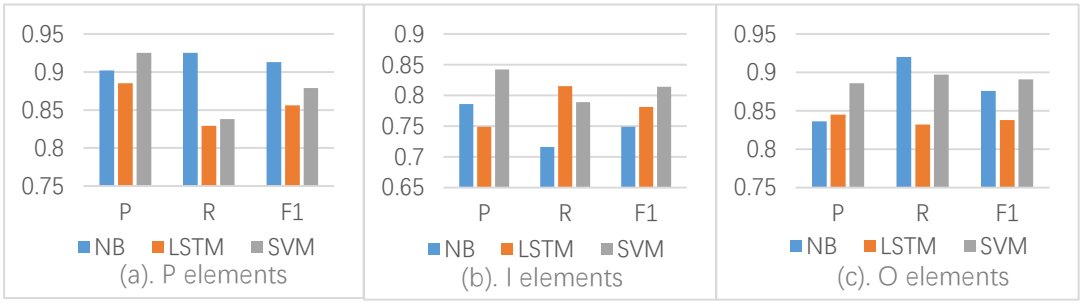

Fig 3: Histograms comparing the experimental results of the SVM model presented in this study, the LSTM model proposed by Di Jin 2018, and the NB model proposed by Ke-Chun Huang 2011: (a) results for P elements, (b) results for I elements, and (c) results for O elements. The three evaluation indicators, i.e., the P, R and F1 values, are represented on the horizontal axes.

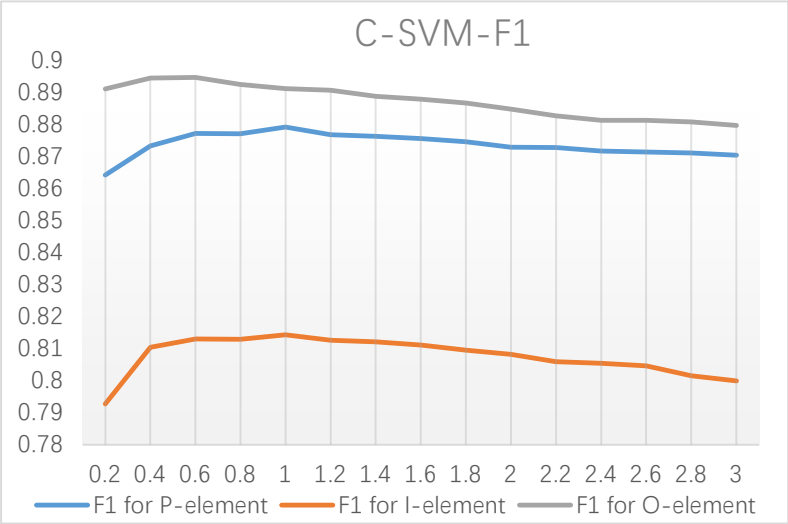

Fig 4: Comparative experiment of the penalty parameter C in SVM model. We found that for P and I elements, the model achieved the best F1 values with C=1.0, whereas for O elements, the model achieved the best F1 value with C=0.6.