# Domain-Adapted Measurement Extraction Using RoBERTa

**Sam Stephens**
**Michelle Shen**
**Matt Lyons**
U.C. Berkeley School of Information
{sam.stephens, michelleshen, mattslyons}@berkeley.edu

## Abstract

SemEval 2021 Task 8: MeasEval: Counts and Measurements aimed to generate research on the contextual extraction of counts and measurements in order to introduce a system of measurement extraction that identified entities while preserving semantic relationships. MeasEval includedfive subtasks ranging from straightforward quantity extraction to more complicated qualifier extractions. Because these tasks primarily concern scientific texts, we completed several MeasEval tasks using base RoBERTa as our baseline, and cs_roberta_base and biomed_roberta_base as competitor models, hypothesizing that the domain-adapted pretrained (DAPT) RoBERTa models, especially the model trained on biomedical texts, would yield improved performance. We demonstrate that DAPT models offer improvement over non-DAPT models even when unique labeled training tokens are repeated as few as two or three times within a domain.

## 1 Introduction

Measurement is at the heart of science. If something is to be understood, it must be examined and its attributes quantified. When researchers communicate quantities, they are supplying the very link between the theory and the real world: *this* is the measurement, and *this* is what it means. But absent the context which gives it meaning, a number is no more concrete than a theory. This is the motivation behind various ontologies of units of measure developed by Rijgersberg et al., 2011 (among others) and surveyed in Steinberg et al., 2017.

Even when conforming to standards of scientific writing, researchers can employ various approaches to presenting their results. Units can precede or follow quantities, which may themselves not be colocated with (or indeed, may be located far from) the entities or properties being measured. Though no doubt an important task, the extraction of these measurements along with related context is also a challenging one, and has only recently seen substantial progress (Hundman and Mattmann, 2017).

Based on one such ontology of units of measure, SemEval-2021 Task 8: MeasEval: Counts and Measurements (Harper et al., 2021) aimed to generate research which extracted not only such quantities from scientific documents, but also the context tying them to the real world: What property of what entity is being measured? On what scale is it being measured? Under what conditions?

MeasEval consisted of five subtasks. Broadly, subtasks 1–4 asked participants to determine the spans corresponding to Quantity, Unit of Measurement, Measured Entity, Measured Property, and Qualifier, where present, in 448 English-language scientific texts provided by the task organizers. Task 5 was to determine the relationship between each of the identified entity types within the passage.

We present a comparison of RoBERTa-base, cs_roberta_base, and biomed_roberta_base models on the first four of these tasks, with the fifth still underway, in order to determine whether the domain-specific improvements described by Gururangan et al. extend to the MeasEval task.

## 2 Background

Although the approaches employed by the two best-performing teams in the competition differed in many respects, both LIORI (Davletov et al., 2021) and jarvis@tencent (Cao et al., 2021) utilized the base model of RoBERTa to classify quantity spans and extract relation information. Based on the tremendously important work of Devlin et al., 2018, who developed the transformer-based language model BERT, RoBERTa (Liu et al., 2019) used a much larger corpus, altered hyperparameters, and made changes to training tasks and procedures to significantly improve performance.

Just as Liu et al. set out to improve BERT's

performance by varying the training process, Gururangan et al., 2020 showed that applying domain-adaptive pretraining (DAPT) to base RoBERTa improved performance on domain-specific tasks. Two outputs from this research were the cs_roberta_base and biomed_roberta_base models, made available by the Allen Institute for Artificial Intelligence, which applied DAPT to base RoBERTa using computer science and biomedical texts.

Recognizing that scientific texts comprise, if not all, then certainly the bulk of the documents in the MeasEval task dataset, our team hypothesized that biomed_roberta_base would perform better than base RoBERTa due to the prevalence of measured quantities in the biomedical sphere, while performance using cs_roberta_base would be the same as or worse than performance using base RoBERTa.

## 3 Methods

We will begin by discussing the format and processing applied to the task-supplied dataset, then move to discussions of the model structure and training process.

### 3.1 Data Format and Preparation

The data provided by the task organizers consisted of 448 English-language scientific texts and their 5281 accompanying annotations, which were supplied for each document as tab-separated files. The provided data were already split into training, development, and test sets, so we retained these splits for consistency with other participant groups.

Within each document, annotations referring to the same quantity were gathered in annotation sets, which collectively supplied all of the location (provided as character indices for the associated text) and type (e.g. Quantity, Measured Property, Measured Entity, Qualifier) information for that annotation. Unit data were included as plain-text within an "other" column associated with each quantity annotation.

We tokenized text using the RoBERTa tokenizer, which is shared by cs_roberta_base and biomed_roberta_base. Unlike some MeasEval submissions which split documents into sentences for training, we left each document intact. We felt that exposing the model to the larger context of each document was important and that splitting by sentence would put too much weight on those sentences contained within longer documents. Documents longer than the 512 token RoBERTa limit were truncated.

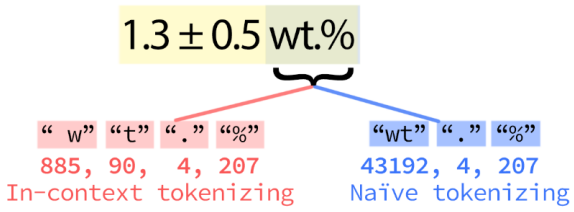


Figure 1: Example of different tokenizations produced by in-context and standalone tokenizing of units

Although some groups (e.g. LIORI) chose to train one model per task, such that one model would aim to identify only measured entities, we felt that a multi-class model would be better-suited to the task of extracting the necessary context around the quantities to relate these labels to one another.

One drawback to this approach was the necessity of each position within the text to have one and only one label. Upon examining the data, we discovered that several dozen spans had conflicting labels between annotation sets. Because we wanted to train our model only on complete annotation sets, we made the decision to bypass annotation sets which would otherwise overwrite spans to which another label had already been applied.

We expected these "collisions" would occur in patterns. For example, perhaps the temperature or pressure at which the volume of a substance was measured, though quantities themselves, served as qualifiers in the quantity annotation representing the volume. This would be represented in the labels by a 4 attempting to overwrite a 1 or the reverse. However, while testing our collision detection, we found that every combination, in every order, of label collisions did indeed occur within the provided data.

An additional complication was how to predict spans for Units of Measurement. On our first attempt to label Unit spans, we tokenized the raw text representing the Unit and searched for matching token sequences within the Quantity span. This would prove problematic in two ways. First, we encountered cases where the Unit was located outside of the Quantity span. To handle these cases, we introduced a margin around the Quantity to search within. Second, we found that a naive tokenization of the Unit text would sometimes produce different tokenizations than the same text encoded in the original sequence (Fig. 1). We solved the issue of


Figure 2: Example of text with supplied annotations. Note that this Unit overlaps the Quantity, so the 5 label has been allowed to overwrite the 1

token alignment by using the `token_to_chars` method to return the index of the first character in the original string associated with the indicated token.

In order to standardize the way the annotations were handled and to streamline preprocessing, we took the novel approach of de-embedding the Units from within the Quantity annotations and treating them as their own labels (unit label=5) to be predicted alongside Quantity, Measured Entity, Measured Property, and Qualifier. This required modifying the collision handling to allow collisions to occur only if a Unit was attempting to overwrite a Quantity within the same annotation set (Fig. 2).

Removing these collisions had the consequence of reducing the total count of annotations by 261, resulting in a total annotation count of 5020 within the corpus of 448 documents. Although this came at a slight cost, we felt that training the labels together would produce better results in terms of the context surrounding the Quantities, and within the constraints imposed by that decision, we chose to remove colliding annotation sets in order to maintain clean and consistent labels.

### 3.2 Training, Evaluation, and Modeling Decisions

The MeasEval task organizers and participants used F1-score as an evaluation metric in order to balance false positives and negatives, so we likewise chose F1-score as our evaluation metric. After completing preprocessing on a combined data set, we split the data back into the training, dev, and test sets specified by the task organizers. The training set was divided into batches of 8 and sent to the device for training.

We built our model around the RoBERTa structure, which has 12 layers with 12 attention heads each and uses a 768-dimensional embedding space. We used the RoBERTa-default GELU activation function. After these 12 layers, the outputs are passed into a pooling layer using tanh activation. On top of this layer, we added a torch.nn.Linear layer which would transform the data into raw prediction values corresponding to each of our 6 classes (Quantity, Entity, Property, Qualifier, Unit, and No Label).

To determine our training loss, we used the sum of the cross entropy loss across these six labels $\sum_{c=1}^{6} y_{o,c} \log(p_{o,c})$. We initialized the learning rate, then used the AdamW optimizer (Loshchilov and Hutter, 2017) to adjust the learning rate as needed during training.

We used random search to optimize the initial learning rate and the dropout rate. Figure 3 shows the search space of hyperparameters and the combinations we evaluated. Prior to hyperparameter tuning, we ran several pipeline diagnostic models for debugging purposes and to explore the hyperparameter search space. To determine the best set of hyperparameters, we summed F1-score evaluated on the development dataset, weighted by task, across each epoch, ultimately selecting a dropout of 0.022 and initial learning rate of 7e-05. Figure 4 shows the F1-score of the winning (and losing) hyperparameter combinations across the training epochs.

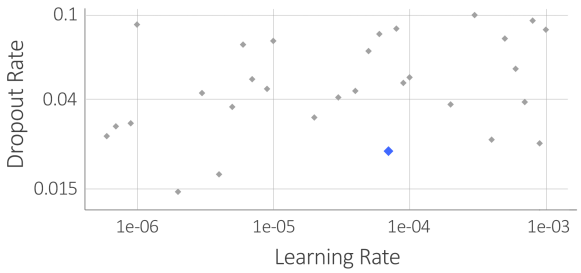


Figure 3: Hyperparameter search space and combinations evaluated; the best combination is shown in blue.

Using these hyperparameters, we trained a baseline model using `roberta-base` as well as our candidate models using `biomed_roberta_base` and `cs_roberta_base`. We trained each model for eight epochs, although dev set performance tended to level off after about three epochs.

## 4 Results and Discussion

A summary of our results, categorized by text domain, is shown in Table 1 and in Figure 6.

Our results are mixed: although the two domain-adaptive pretrained models performed better in a handful of areas, the RoBERTa base model saw better performance across several domains. Computer

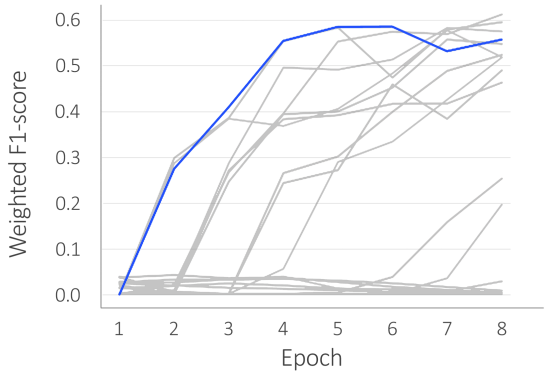


Figure 4: F1-scores by epoch during training for various hyperparameter combinations. The combination with the greatest sum across all epochs (shows in blue) was selected for use.

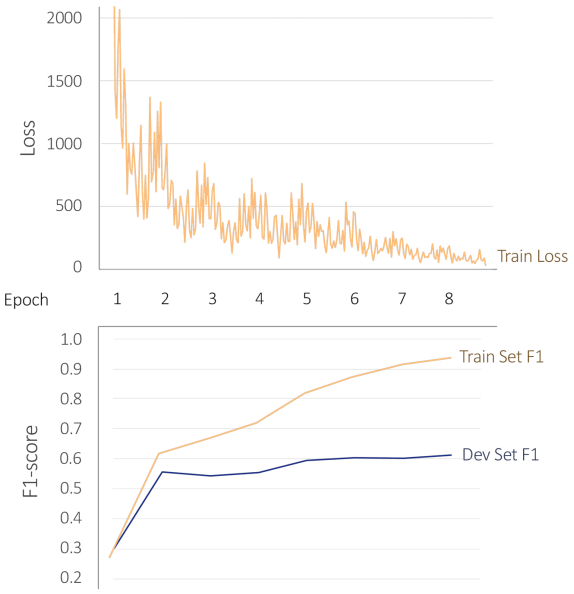


Figure 5: Loss during training and F1-score evaluated per epoch for training and dev sets. Dev set F1-score levels off after epoch 2, while training loss continues to decrease.

science trained RoBERTa performed the worst overall, with an average F1-score of 0.592. It only outperformed `roberta-base` on biology texts. The `biomed` model performed slightly better, outperforming `roberta-base` in texts on agriculture, astronomy, biology, computer science, and medicine and earning an overall F1-score of 0.617. However, the baseline model performed best in chemistry, earth science, engineering, materials science, and mathematics, with an overall F1-score of 0.641. We note that there was only one mathematics text in the test set provided by the task organizers, which helps to explain the notable performance difference.

### 4.1 Token-Level Analysis

In order to better understand our results, we looked at the types of tokens each model got right and wrong most frequently and organized them into broad categories. A table of ranked tokens by frequency of misprediction for each model is available in the appendix.

Broadly, the most frequently mispredicted tokens were words or symbols that sometimes represent a relationship to a quantity, but which also occur in non-numeric contexts. For example, a hyphen may function as a minus sign used for subtraction, denote separate ranges of numbers, or act as a separator for words unrelated to quantities. A period or en-dash are similarly ambiguous. Symbols and punctuation marks like these top the list of mispredicted spans for each of the three model types.

Following punctuation, prepositions were the most frequently mispredicted tokens for the roberta-base and `biomed` models. As above, these are words that often relate to numbers or quantities, but just as often may not. Words like “of”, “from”, “over”, “per”, and “for” are all often used in contexts that may or may not relate to measurements–or which may relate to different types of measurements in different ways. It is common in some NLP tasks to count some of these words as “stop words” (words which occur so frequently so as to represent noise more than they do signal) and remove them. However, in the context of quantities and measurements, many prepositions like these fundamentally alter the contextual meaning of the numbers near which they occur, so removing them would not have been a realistic option. In lieu of the normal stop words approach, perhaps one way to improve performance on this task would be to carefully construct a list of stop words and symbols which do not ordinarily interact with or change the meaning of quantities and filter out tokens contained in this list.

Interestingly, unlike with the `roberta-base` and `biomed` models, the second most frequently mispredicted category of tokens for the `cs` model were single letters (or letter combinations with whitespace), followed by prepositions. We hypothesize that this is the case because variable names and equations (in these cases *not* referring to quantities or measurements) are more frequently used

| Domain | roberta-base | biomed | % change | cs | % change |
|---|---|---|---|---|---|
| **Agriculture** | 0.607 | **0.616** | 0.015 | 0.588 | -0.031 |
| **Astronomy** | 0.633 | **0.640** | 0.012 | 0.595 | -0.059 |
| **Biology** | 0.595 | 0.609 | 0.023 | **0.612** | 0.028 |
| **Chemistry** | **0.721** | 0.718 | -0.003 | 0.708 | -0.017 |
| **Computer Science** | 0.474 | **0.517** | 0.093 | 0.432 | -0.087 |
| **Earth Science** | **0.560** | 0.549 | -0.021 | 0.520 | -0.072 |
| **Engineering** | **0.612** | 0.573 | -0.064 | 0.577 | -0.057 |
| **Materials Science** | **0.638** | 0.617 | -0.034 | 0.578 | -0.094 |
| **Mathematics** | **0.933** | 0.667 | -0.286 | 0.711 | -0.238 |
| **Medicine** | 0.633 | **0.661** | 0.044 | 0.595 | -0.059 |
| **Average** | **0.641** | **0.617** | **-0.037** | **0.592** | **-0.076** |

Table 1: Overall Performance Comparison of roberta-base, biomed, and cs (F1-scores)

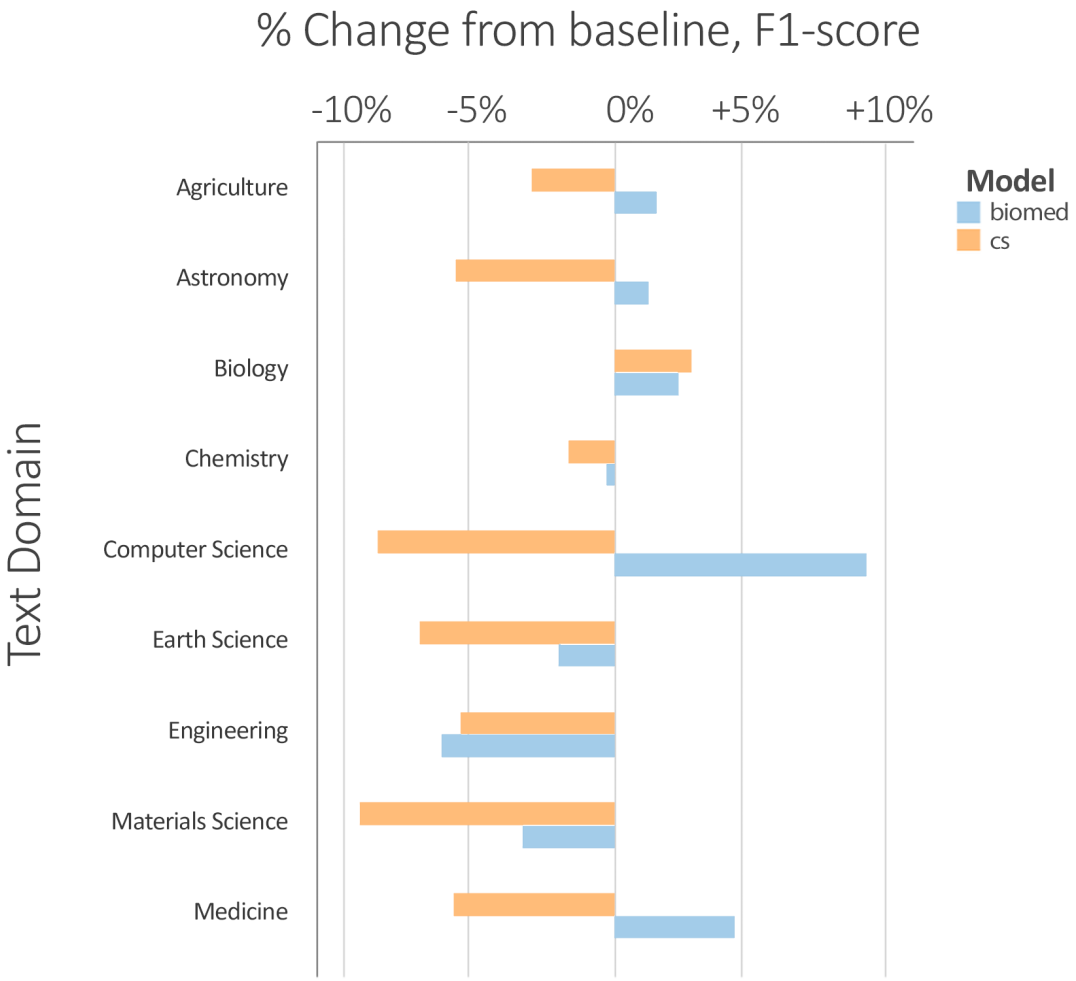


Figure 6: DAPT Model Performance vs `roberta-base`

in computer science writing than in the natural sciences. Likewise, we suspect that units commonly seen in other fields (e.g. m as in meters, C as in degrees Celsius, g as in grams) may occur rarely, if at all, within computer science texts.

Unsurprisingly, the fourth most frequently mispredicted token category for each model type was numbers (some Arabic numerals and some spelled out, e.g. "two"). As with the categories mentioned so far, though they often refer to quantities or measurements, numbers are also frequently used in other contexts (e.g. to denote years or as figure or table numbers). As part of their preprocessing, some MeasEval participant groups chose to convert all numbers to 0s. It was our intuition that retaining numbers as presented in the texts, for the same reasons that we used a multiclass model instead of one model per subtask, would better preserve the meaning and context of the sentence as a whole, but perhaps this was not the case. This may also be an issue of undertraining: a model may not have had a large enough corpus of documents to encounter the many meanings and contexts numbers may take on and appear in.

In fifth place for most frequently mispredicted category for all three model types were domain-specific words and acronyms such as "flux," "electron," "soil," and "MRI." We expected that a model's exposure to words in this category could be limited and inconsistent.

### 4.2 Impact of Repeated Training Tokens

We also hypothesized that the training corpus did not contain sufficient repeated labeled tokens to train the model on the various subtasks. To validate this hypothesis, we examined the frequency at which unique tokens were labeled (see Table 2). Only tokens labeled as Quantity or Unit were labeled at a rate more than twice per unique token in the training set. These low repeat rates in training labels offer an explanation for the by-domain performance seen in the confusion matrices in Figure 7. Note the high normalized accuracy for Quantity and Unit and low scores for all other labeled subtasks.

Limiting our analysis to the Quantity and Unit subtasks, the DAPT models did demonstrate increased F1 performance over the non-DAPT model in some domain and subtask combinations, but suffered in others (Figure 8). Unit subtask performance improvement was as expected with biomedical DAPT demonstrating increased unit classification performance in the domains of Biology, Medicine, Chemistry, and Agriculture; which we consider biomedical-related domains. Computer science DAPT also demonstrated unit classification performance on that domain. This strong performance may be explained by repeat rate of Unit labeled unique tokens, which ranged from about 3 to 4 in each domain.

| Label Name | Labels per Unique Token |
|---|---|
| NoLabel | 19.248 |
| Quantity | 5.0439 |
| MeasuredProperty | **1.725** |
| MeasuredEntity | **1.989** |
| Qualifier | **1.577** |
| Unit | 5.473 |

Table 2: Frequency analysis of labels per unique token. Note how few labels were present each for Measured Property, Measured Entity, and Qualifier.

DAPT performance gains were also seen with Quantity classification, but not as clearly as with Unit. Computer Science and Medicine were the two domains with no DAPT advantage on Quantity classification. The lack of improvement in the Computer Science domain with CS DAPT can be explained by an exceptionally low repeat rate of labeled Quantity tokens in that domain of 1.81. The biomedical DAPT lack of improvement in the Medical domain may be explained similarly, but that conclusion is less clear with a Medical domain repeat token label rate of 2.06.

## 5 Conclusion

In most domains and subtasks where training set labeled tokens were repeated at least twice the DAPT models demonstrated improved classification performance over the non-DAPT model. This was in line with our hypothesis. We saw no consistent domain-specific performance gains from the DAPT models in subtasks where labeled tokens were repeated less often. However, it is notable that DAPT models are able to show improvement on these subtasks in domains with fewer than 5 repeated labeled tokens in the training set. Our recommendation is that DAPT RoBERTa models for MeasEval classification subtasks be trained on a corpus where unique subtask tokens occur at least 3 times per domain in varying contexts.

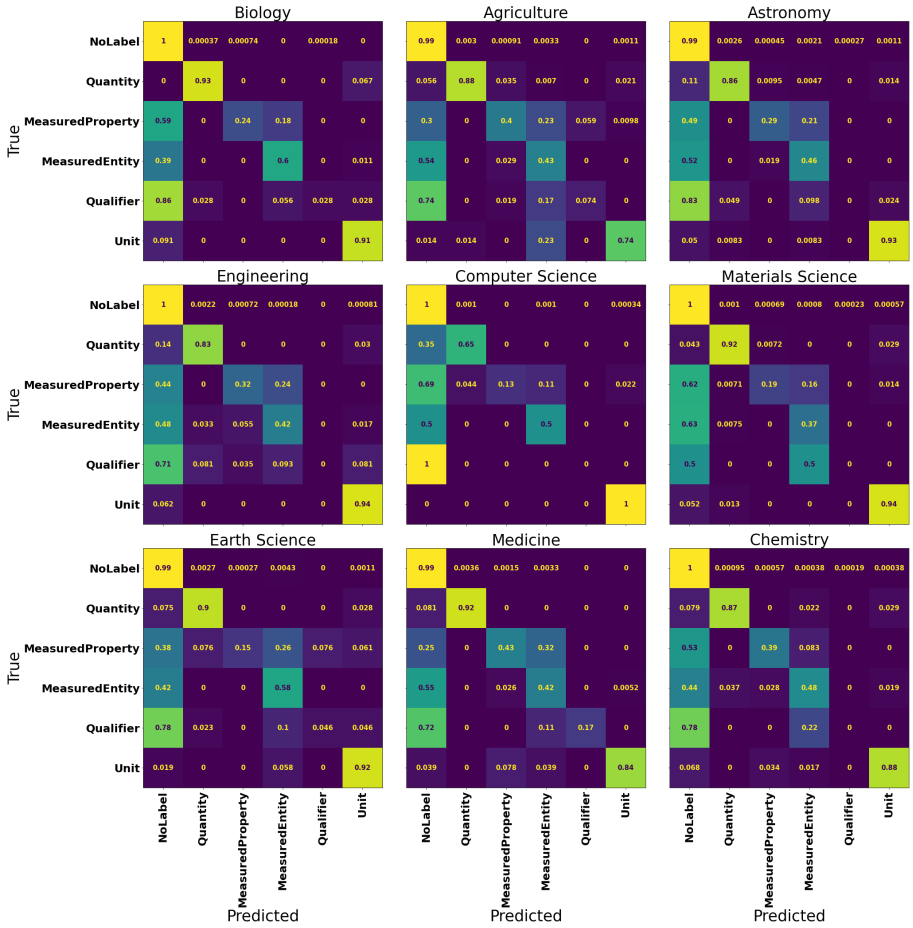


Figure 7: Confusion matrices for each task within each domain

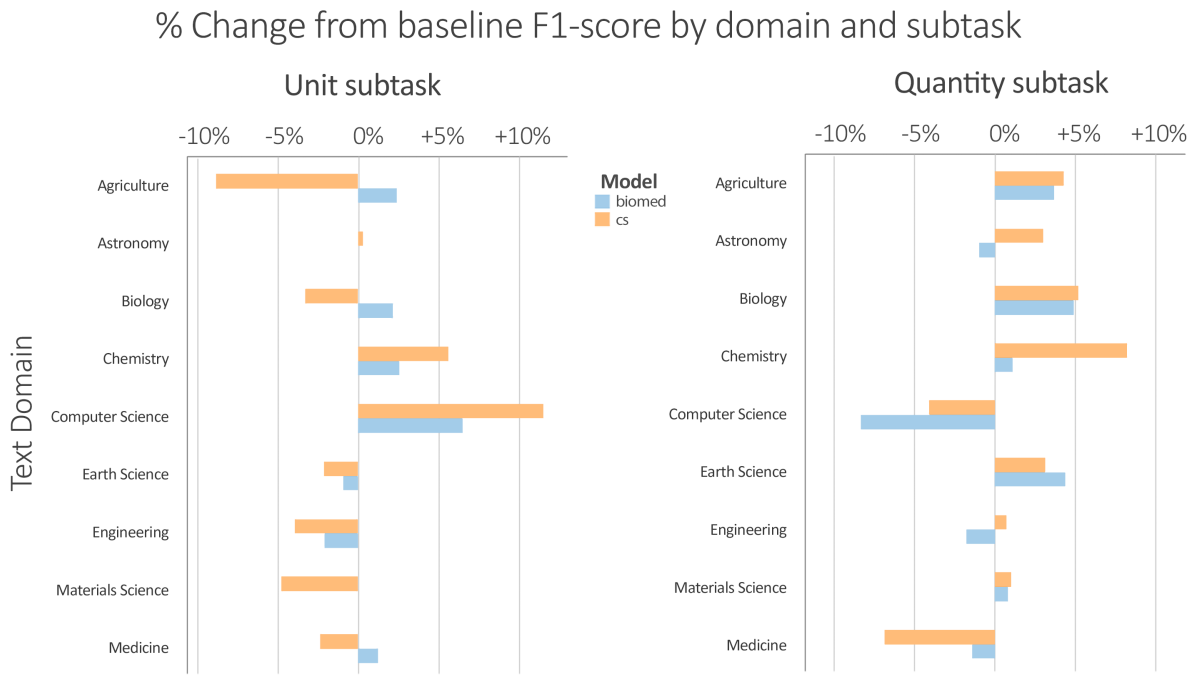


Figure 8: Percent change in F1-score from `roberta-base` model for the Unit and Quantity subtasks. Other subtasks with low labeled token repeat rates showed no consistent improvement pattern and are not displayed.

## References


Jiarun Cao, Yuejia Xiang, Yunyan Zhang, Zhiyuan Qi, Xi Chen, and Yefeng Zheng. 2021. CONNER: A cascade count and measurement extraction tool for scientific discourse. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1239–1244, Online. Association for Computational Linguistics.

Adis Davletov, Denis Gordeev, Nikolay Arefyev, and Emil Davletov. 2021. LIORI at SemEval-2021 task 8: Ask transformer for measurements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1249–1254, Online. Association for Computational Linguistics.

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

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *CoRR*, abs/2004.10964.

Corey Harper, Jessica Cox, Curt Kohler, Antony Scerri, Ron Daniel Jr., and Paul Groth. 2021. SemEval-2021 task 8: MeasEval – extracting counts and measurements and their related contexts. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 306–316, Online. Association for Computational Linguistics.

Kyle Hundman and Chris A. Mattmann. 2017. Measurement context extraction from text: Discovering opportunities and gaps in earth science.

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.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization.

H. Rijgersberg, M. Wigham, and J.L. Top. 2011. How semantics can improve engineering processes: A case of units of measure and quantities. *Advanced Engineering Informatics*, 25(2):276–287. Information mining and retrieval in design.

Markus D. Steinberg, Sirko Schindler, and Jan Martin Keil. 2017. Use cases and suitability metrics for unit ontologies. In *OWL: Experiences and Directions – Reasoner Evaluation*, pages 40–54, Cham. Springer International Publishing.

# 1 Appendix A: Table of Mispredicted Tokens and Frequencies

Presented here are the tokens most frequently mispredicted by each model on the test set, along with a categorization and the count of mispredictions for that category for that model.

## 1.1 Mispredictions of the `roberta-base` model

Note: some symbols are not displayed due to encoding issues.

| Token Text | Token Count | Category | Category Count |
|---|---|---|---|
| - | 55 | punct/symbol | 256 |
| | 40 | punct/symbol | 256 |
| . | 32 | punct/symbol | 256 |
| , | 22 | punct/symbol | 256 |
| ) | 21 | punct/symbol | 256 |
| ( | 19 | punct/symbol | 256 |
| | 16 | punct/symbol | 256 |
| ( | 7 | punct/symbol | 256 |
| ° | 6 | punct/symbol | 256 |
| | 5 | punct/symbol | 256 |
| ± | 4 | punct/symbol | 256 |
| / | 4 | punct/symbol | 256 |
| = | 3 | punct/symbol | 256 |
| | 3 | punct/symbol | 256 |
| × | 3 | punct/symbol | 256 |
| of | 36 | preposition | 167 |
| in | 24 | preposition | 167 |
| to | 17 | preposition | 167 |
| from | 11 | preposition | 167 |
| at | 9 | preposition | 167 |
| than | 9 | preposition | 167 |
| for | 7 | preposition | 167 |
| over | 6 | preposition | 167 |
| per | 6 | preposition | 167 |
| in | 6 | preposition | 167 |
| within | 5 | preposition | 167 |
| above | 4 | preposition | 167 |
| as | 4 | preposition | 167 |
| between | 4 | preposition | 167 |
| on | 4 | preposition | 167 |
| below | 3 | preposition | 167 |
| near | 3 | preposition | 167 |
| with | 3 | preposition | 167 |
| Between | 3 | preposition | 167 |
| on | 3 | preposition | 167 |
| C | 16 | letter | 144 |

Continued on next page

**Table 1 – `base` – continued from previous page**

| Token Text | Token Count | Category | Category Count |
|---|---|---|---|
| m | 13 | letter | 144 |
| G | 9 | letter | 144 |
| C | 7 | letter | 144 |
| S | 7 | letter | 144 |
| E | 6 | letter | 144 |
| c | 6 | letter | 144 |
| x | 6 | letter | 144 |
| R | 6 | letter | 144 |
| R | 5 | letter | 144 |
| s | 5 | letter | 144 |
| N | 5 | letter | 144 |
| t | 5 | letter | 144 |
| S | 4 | letter | 144 |
| k | 4 | letter | 144 |
| Z | 4 | letter | 144 |
| H | 3 | letter | 144 |
| K | 3 | letter | 144 |
| O | 3 | letter | 144 |
| f | 3 | letter | 144 |
| g | 3 | letter | 144 |
| p | 3 | letter | 144 |
| E | 3 | letter | 144 |
| M | 3 | letter | 144 |
| a | 3 | letter | 144 |
| d | 3 | letter | 144 |
| p | 3 | letter | 144 |
| s | 3 | letter | 144 |
| 2 | 19 | number | 127 |
| two | 17 | number | 127 |
| 1 | 9 | number | 127 |
| 5 | 9 | number | 127 |
| 2 | 7 | number | 127 |
| 4 | 6 | number | 127 |
| 1 | 5 | number | 127 |
| 12 | 5 | number | 127 |
| 13 | 5 | number | 127 |
| 3 | 5 | number | 127 |
| 0 | 4 | number | 127 |
| 13 | 4 | number | 127 |
| 20 | 4 | number | 127 |
| 4 | 4 | number | 127 |
| three | 4 | number | 127 |
| 6 | 4 | number | 127 |
| 7 | 4 | number | 127 |
| 3 | 3 | number | 127 |
| 8 | 3 | number | 127 |

Continued on next page

**Table 1 – `base` – continued from previous page**

| Token Text | Token Count | Category | Category Count |
|---|---|---|---|
| 0 | 3 | number | 127 |
| 10 | 3 | number | 127 |
| flux | 7 | noun | 96 |
| activity | 6 | noun | 96 |
| layer | 6 | noun | 96 |
| model | 6 | noun | 96 |
| conditions | 5 | noun | 96 |
| electron | 5 | noun | 96 |
| energy | 5 | noun | 96 |
| algorithms | 4 | noun | 96 |
| heating | 4 | noun | 96 |
| cells | 3 | noun | 96 |
| class | 3 | noun | 96 |
| differences | 3 | noun | 96 |
| electrons | 3 | noun | 96 |
| groups | 3 | noun | 96 |
| injection | 3 | noun | 96 |
| lid | 3 | noun | 96 |
| reaction | 3 | noun | 96 |
| setting | 3 | noun | 96 |
| species | 3 | noun | 96 |
| speeds | 3 | noun | 96 |
| turbine | 3 | noun | 96 |
| wind | 3 | noun | 96 |
| member | 3 | noun | 96 |
| end | 3 | noun | 96 |
| time | 3 | noun | 96 |
| the | 51 | article | 64 |
| a | 9 | article | 64 |
| A | 4 | article | 64 |
| average | 7 | quantity word | 51 |
| approximately | 6 | quantity word | 51 |
| more | 5 | quantity word | 51 |
| rate | 5 | quantity word | 51 |
| lower | 4 | quantity word | 51 |
| factor | 3 | quantity word | 51 |
| increase | 3 | quantity word | 51 |
| less | 3 | quantity word | 51 |
| output | 3 | quantity word | 51 |
| reach | 3 | quantity word | 51 |
| total | 3 | quantity word | 51 |
| upper | 3 | quantity word | 51 |
| values | 3 | quantity word | 51 |
| dual | 4 | adjective | 35 |
| not | 4 | adjective | 35 |
| positive | 4 | adjective | 35 |

Continued on next page

**Table 1 – `base` – continued from previous page**

| Token Text | Token Count | Category | Category Count |
|---|---|---|---|
| rough | 4 | adjective | 35 |
| old | 4 | adjective | 35 |
| auditory | 3 | adjective | 35 |
| both | 3 | adjective | 35 |
| each | 3 | adjective | 35 |
| long | 3 | adjective | 35 |
| other | 3 | adjective | 35 |
| osp | 5 | short letter combo | 34 |
| ES | 4 | short letter combo | 34 |
| yl | 4 | short letter combo | 34 |
| ST | 3 | short letter combo | 34 |
| Str | 3 | short letter combo | 34 |
| HC | 3 | short letter combo | 34 |
| MRI | 3 | short letter combo | 34 |
| SF | 3 | short letter combo | 34 |
| VD | 3 | short letter combo | 34 |
| ch | 3 | short letter combo | 34 |
| ion | 5 | suffix | 33 |
| inter | 4 | suffix | 33 |
| 's | 4 | suffix | 33 |
| ar | 4 | suffix | 33 |
| heric | 4 | suffix | 33 |
| ing | 3 | suffix | 33 |
| ness | 3 | suffix | 33 |
| osphere | 3 | suffix | 33 |
| ulus | 3 | suffix | 33 |
| and | 28 | conjunction | 28 |
| temperature | 5 | measure | 11 |
| density | 3 | measure | 11 |
| resolution | 3 | measure | 11 |
| mm | 5 | unit | 8 |
| kg | 3 | unit | 8 |
| CO | 7 | chemical | 7 |
| surface | 6 | location | 6 |
| CDC | 4 | acronym | 4 |
| did | 3 | aux verb | 3 |
| that | 3 | stop word | 3 |
| amber | 3 | item | 3 |
| derived | 3 | verb | 3 |

## 1.2 Mispredictions of the `biomed_roberta_base` model

Note: some symbols are not displayed due to encoding issues.

| Token Text | Token Count | Category | Category Count |
|---|---|---|---|
| | 44 | punct/symbol | 236 |
| - | 44 | punct/symbol | 236 |
| . | 27 | punct/symbol | 236 |
| , | 20 | punct/symbol | 236 |
| ) | 18 | punct/symbol | 236 |
| | 15 | punct/symbol | 236 |
| ( | 15 | punct/symbol | 236 |
| ° | 6 | punct/symbol | 236 |
| ( | 6 | punct/symbol | 236 |
| / | 5 | punct/symbol | 236 |
| | 3 | punct/symbol | 236 |
| = | 3 | punct/symbol | 236 |
| ¡ | 3 | punct/symbol | 236 |
| × | 3 | punct/symbol | 236 |
| – | 3 | punct/symbol | 236 |
| | 3 | punct/symbol | 236 |
| ± | 3 | punct/symbol | 236 |
| of | 31 | preposition | 157 |
| in | 22 | preposition | 157 |
| to | 21 | preposition | 157 |
| from | 19 | preposition | 157 |
| at | 9 | preposition | 157 |
| in | 8 | preposition | 157 |
| for | 7 | preposition | 157 |
| than | 7 | preposition | 157 |
| per | 6 | preposition | 157 |
| on | 4 | preposition | 157 |
| with | 4 | preposition | 157 |
| over | 4 | preposition | 157 |
| within | 3 | preposition | 157 |
| on | 3 | preposition | 157 |
| between | 3 | preposition | 157 |
| below | 3 | preposition | 157 |
| as | 3 | preposition | 157 |
| C | 15 | letter | 144 |
| m | 14 | letter | 144 |
| R | 8 | letter | 144 |
| C | 8 | letter | 144 |
| x | 6 | letter | 144 |
| G | 6 | letter | 144 |
| t | 6 | letter | 144 |
| k | 5 | letter | 144 |
| N | 5 | letter | 144 |

Continued on next page

**Table 2 – `biomed` – continued from previous page**

| Token Text | Token Count | Category | Category Count |
|---|---|---|---|
| S | 5 | letter | 144 |
| R | 4 | letter | 144 |
| T | 4 | letter | 144 |
| r | 4 | letter | 144 |
| O | 4 | letter | 144 |
| f | 4 | letter | 144 |
| E | 4 | letter | 144 |
| c | 4 | letter | 144 |
| V | 4 | letter | 144 |
| g | 4 | letter | 144 |
| d | 3 | letter | 144 |
| b | 3 | letter | 144 |
| p | 3 | letter | 144 |
| s | 3 | letter | 144 |
| i | 3 | letter | 144 |
| M | 3 | letter | 144 |
| L | 3 | letter | 144 |
| E | 3 | letter | 144 |
| S | 3 | letter | 144 |
| K | 3 | letter | 144 |
| 2 | 17 | number | 134 |
| two | 15 | number | 134 |
| 5 | 10 | number | 134 |
| 1 | 8 | number | 134 |
| 1 | 7 | number | 134 |
| 13 | 6 | number | 134 |
| 2 | 6 | number | 134 |
| three | 5 | number | 134 |
| 10 | 5 | number | 134 |
| 3 | 5 | number | 134 |
| 4 | 4 | number | 134 |
| 20 | 4 | number | 134 |
| 4 | 4 | number | 134 |
| 0 | 4 | number | 134 |
| 12 | 4 | number | 134 |
| 23 | 3 | number | 134 |
| 10 | 3 | number | 134 |
| 6 | 3 | number | 134 |
| 13 | 3 | number | 134 |
| 8 | 3 | number | 134 |
| 60 | 3 | number | 134 |
| 5 | 3 | number | 134 |
| 3 | 3 | number | 134 |
| 7 | 3 | number | 134 |
| 0 | 3 | number | 134 |
| activity | 7 | noun | 87 |

Continued on next page

**Table 2 – `biomed` – continued from previous page**

| Token Text | Token Count | Category | Category Count |
|---|---|---|---|
| conditions | 6 | noun | 87 |
| energy | 6 | noun | 87 |
| flux | 6 | noun | 87 |
| layer | 5 | noun | 87 |
| electron | 5 | noun | 87 |
| algorithms | 4 | noun | 87 |
| heating | 4 | noun | 87 |
| class | 4 | noun | 87 |
| cells | 4 | noun | 87 |
| electrons | 3 | noun | 87 |
| groups | 3 | noun | 87 |
| sand | 3 | noun | 87 |
| model | 3 | noun | 87 |
| magnet | 3 | noun | 87 |
| differences | 3 | noun | 87 |
| lid | 3 | noun | 87 |
| species | 3 | noun | 87 |
| setting | 3 | noun | 87 |
| years | 3 | noun | 87 |
| September | 3 | noun | 87 |
| end | 3 | noun | 87 |
| the | 49 | article | 54 |
| a | 5 | article | 54 |
| osp | 4 | short letter combo | 38 |
| yl | 4 | short letter combo | 38 |
| ch | 3 | short letter combo | 38 |
| VD | 3 | short letter combo | 38 |
| ES | 3 | short letter combo | 38 |
| MS | 3 | short letter combo | 38 |
| AE | 3 | short letter combo | 38 |
| MRI | 3 | short letter combo | 38 |
| ST | 3 | short letter combo | 38 |
| PS | 3 | short letter combo | 38 |
| OC | 3 | short letter combo | 38 |
| SOC | 3 | short letter combo | 38 |
| auditory | 6 | adjective | 36 |
| rough | 4 | adjective | 36 |
| not | 4 | adjective | 36 |
| rough | 4 | adjective | 36 |
| both | 3 | adjective | 36 |
| other | 3 | adjective | 36 |
| neutral | 3 | adjective | 36 |
| positive | 3 | adjective | 36 |
| transient | 3 | adjective | 36 |
| dual | 3 | adjective | 36 |
| average | 7 | quantity word | 30 |

Continued on next page

**Table 2 – `biomed` – continued from previous page**

| Token Text | Token Count | Category | Category Count |
|---|---|---|---|
| lower | 5 | quantity word | 30 |
| rate | 5 | quantity word | 30 |
| more | 4 | quantity word | 30 |
| factor | 3 | quantity word | 30 |
| values | 3 | quantity word | 30 |
| increase | 3 | quantity word | 30 |
| and | 27 | conjunction | 27 |
| heric | 4 | suffix | 22 |
| ar | 3 | suffix | 22 |
| ness | 3 | suffix | 22 |
| ing | 3 | suffix | 22 |
| 's | 3 | suffix | 22 |
| inter | 3 | suffix | 22 |
| ion | 3 | suffix | 22 |
| temperature | 9 | measure | 15 |
| resolution | 3 | measure | 15 |
| power | 3 | measure | 15 |
| surface | 8 | location | 8 |
| CO | 7 | chemical | 7 |
| mm | 4 | unit | 7 |
| kg | 3 | unit | 7 |
| derived | 3 | verb | 6 |
| left | 3 | verb | 6 |
| amber | 3 | item | 3 |
| did | 3 | aux verb | 3 |
| that | 3 | stop word | 3 |

### 1.3 Mispredictions of the `cs_roberta_base` model

Note: some symbols are not displayed due to encoding issues.

| Token Text | Token Count | Category | Category Count |
|---|---|---|---|
| - | 43 | punct/symbol | 236 |
| | 37 | punct/symbol | 236 |
| . | 28 | punct/symbol | 236 |
| , | 20 | punct/symbol | 236 |
| ) | 17 | punct/symbol | 236 |
| ( | 15 | punct/symbol | 236 |
| | 13 | punct/symbol | 236 |
| | 7 | punct/symbol | 236 |
| ( | 6 | punct/symbol | 236 |
| / | 6 | punct/symbol | 236 |
| ° | 6 | punct/symbol | 236 |
| ¡ | 3 | punct/symbol | 236 |
| ± | 3 | punct/symbol | 236 |
| | 3 | punct/symbol | 236 |
| = | 3 | punct/symbol | 236 |
| – | 3 | punct/symbol | 236 |
| × | 3 | punct/symbol | 236 |
| C | 16 | letter | 166 |
| m | 14 | letter | 166 |
| C | 9 | letter | 166 |
| G | 7 | letter | 166 |
| R | 7 | letter | 166 |
| T | 6 | letter | 166 |
| t | 6 | letter | 166 |
| S | 6 | letter | 166 |
| x | 6 | letter | 166 |
| M | 5 | letter | 166 |
| E | 5 | letter | 166 |
| r | 4 | letter | 166 |
| L | 4 | letter | 166 |
| c | 4 | letter | 166 |
| F | 4 | letter | 166 |
| M | 4 | letter | 166 |
| f | 4 | letter | 166 |
| H | 4 | letter | 166 |
| s | 4 | letter | 166 |
| R | 4 | letter | 166 |
| k | 4 | letter | 166 |
| b | 3 | letter | 166 |
| g | 3 | letter | 166 |
| E | 3 | letter | 166 |
| d | 3 | letter | 166 |
| g | 3 | letter | 166 |

Continued on next page

**Table 3 – `cs` – continued from previous page**

| Token Text | Token Count | Category | Category Count |
|---|---|---|---|
| S | 3 | letter | 166 |
| T | 3 | letter | 166 |
| i | 3 | letter | 166 |
| N | 3 | letter | 166 |
| a | 3 | letter | 166 |
| K | 3 | letter | 166 |
| O | 3 | letter | 166 |
| p | 3 | letter | 166 |
| of | 34 | preposition | 164 |
| in | 23 | preposition | 164 |
| to | 19 | preposition | 164 |
| from | 16 | preposition | 164 |
| at | 9 | preposition | 164 |
| for | 7 | preposition | 164 |
| in | 6 | preposition | 164 |
| than | 6 | preposition | 164 |
| per | 6 | preposition | 164 |
| over | 5 | preposition | 164 |
| above | 4 | preposition | 164 |
| below | 4 | preposition | 164 |
| with | 4 | preposition | 164 |
| within | 4 | preposition | 164 |
| on | 4 | preposition | 164 |
| top | 4 | preposition | 164 |
| as | 3 | preposition | 164 |
| between | 3 | preposition | 164 |
| after | 3 | preposition | 164 |
| 2 | 25 | number | 147 |
| two | 15 | number | 147 |
| 1 | 11 | number | 147 |
| 5 | 10 | number | 147 |
| 1 | 8 | number | 147 |
| 3 | 7 | number | 147 |
| 6 | 6 | number | 147 |
| 2 | 6 | number | 147 |
| 12 | 5 | number | 147 |
| 13 | 5 | number | 147 |
| 4 | 4 | number | 147 |
| 0 | 4 | number | 147 |
| 20 | 4 | number | 147 |
| 4 | 4 | number | 147 |
| 7 | 3 | number | 147 |
| 6 | 3 | number | 147 |
| 2009 | 3 | number | 147 |
| three | 3 | number | 147 |
| 10 | 3 | number | 147 |

Continued on next page

**Table 3 – cs – continued from previous page**

| Token Text | Token Count | Category | Category Count |
|---|---|---|---|
| 209 | 3 | number | 147 |
| 23 | 3 | number | 147 |
| 3 | 3 | number | 147 |
| 458 | 3 | number | 147 |
| 60 | 3 | number | 147 |
| 13 | 3 | number | 147 |
| flux | 7 | noun | 88 |
| energy | 7 | noun | 88 |
| layer | 6 | noun | 88 |
| conditions | 5 | noun | 88 |
| class | 5 | noun | 88 |
| model | 5 | noun | 88 |
| electron | 5 | noun | 88 |
| activity | 4 | noun | 88 |
| electrons | 4 | noun | 88 |
| heating | 4 | noun | 88 |
| groups | 3 | noun | 88 |
| lid | 3 | noun | 88 |
| cells | 3 | noun | 88 |
| injection | 3 | noun | 88 |
| magnet | 3 | noun | 88 |
| soil | 3 | noun | 88 |
| algorithms | 3 | noun | 88 |
| setting | 3 | noun | 88 |
| signatures | 3 | noun | 88 |
| years | 3 | noun | 88 |
| time | 3 | noun | 88 |
| year | 3 | noun | 88 |
| IM | 7 | short letter combo | 66 |
| OM | 6 | short letter combo | 66 |
| osp | 5 | short letter combo | 66 |
| ARS | 5 | short letter combo | 66 |
| yl | 4 | short letter combo | 66 |
| CS | 3 | short letter combo | 66 |
| MS | 3 | short letter combo | 66 |
| ch | 3 | short letter combo | 66 |
| MRI | 3 | short letter combo | 66 |
| PS | 3 | short letter combo | 66 |
| iga | 3 | short letter combo | 66 |
| VD | 3 | short letter combo | 66 |
| Wh | 3 | short letter combo | 66 |
| Str | 3 | short letter combo | 66 |
| th | 3 | short letter combo | 66 |
| ST | 3 | short letter combo | 66 |
| SOC | 3 | short letter combo | 66 |
| ES | 3 | short letter combo | 66 |

Continued on next page

**Table 3 – cs – continued from previous page**

| Token Text | Token Count | Category | Category Count |
|---|---|---|---|
| average | 7 | quantity word | 60 |
| total | 7 | quantity word | 60 |
| mean | 6 | quantity word | 60 |
| lower | 5 | quantity word | 60 |
| rate | 4 | quantity word | 60 |
| peak | 4 | quantity word | 60 |
| factor | 3 | quantity word | 60 |
| more | 3 | quantity word | 60 |
| level | 3 | quantity word | 60 |
| increasing | 3 | quantity word | 60 |
| increase | 3 | quantity word | 60 |
| reach | 3 | quantity word | 60 |
| range | 3 | quantity word | 60 |
| values | 3 | quantity word | 60 |
| variance | 3 | quantity word | 60 |
| the | 49 | article | 55 |
| a | 6 | article | 55 |
| auditory | 7 | adjective | 36 |
| old | 5 | adjective | 36 |
| not | 4 | adjective | 36 |
| dual | 4 | adjective | 36 |
| each | 4 | adjective | 36 |
| annual | 3 | adjective | 36 |
| both | 3 | adjective | 36 |
| other | 3 | adjective | 36 |
| positive | 3 | adjective | 36 |
| temperature | 15 | measure | 34 |
| heric | 5 | suffix | 34 |
| density | 5 | measure | 34 |
| 's | 4 | suffix | 34 |
| precipitation | 4 | measure | 34 |
| power | 4 | measure | 34 |
| ised | 4 | suffix | 34 |
| ing | 4 | suffix | 34 |
| ated | 4 | suffix | 34 |
| ion | 4 | suffix | 34 |
| ar | 3 | suffix | 34 |
| temperatures | 3 | measure | 34 |
| ifier | 3 | suffix | 34 |
| inter | 3 | suffix | 34 |
| pressure | 3 | measure | 34 |
| and | 24 | conjunction | 24 |
| surface | 8 | location | 8 |
| mm | 5 | unit | 8 |
| kg | 3 | unit | 8 |
| CO | 6 | chemical | 6 |

Continued on next page

**Table 3 – `cs` – continued from previous page**

| Token Text | Token Count | Category | Category Count |
|---|---|---|---|
| derived | 3 | verb | 6 |
| left | 3 | verb | 6 |
| did | 3 | aux verb | 3 |
| amber | 3 | item | 3 |
| that | 3 | stop word | 3 |