We extract total of 1,642 unique abstracts, using PubMed id generated by the database. As explained in dataset documentation they generate a document in this format ( Pubmed-ID\n\nTitle\n\nAbstract) Table 1 shows the example of a Document

Table 1 Sample of Document

|  |  |
| --- | --- |
| Pubmed-id | 10048291 |
| Title | Unaccountable severe hypercalcemia in a patient treated for hypoparathyroidism with dihydrotachysterol. |
| Abstract | This report describes a forty-seven-year-old female patient with a complex medical history. She was suffering from an unspecified interstitial lung disease, papillary thyroid carcinoma which had been treated, hypoparathyroidism after thyroidectomy for which she was receiving dihydrotachysterol and calcium, and atrial fibrillation and congestive heart failure as a result of mitral stenosis. … |

Then we normalize the text and split each document to sentences. Each sentence will assign to Adverse effect (AE) of a drug, Dose of a drugs (Dose) or NEG which means that no adverse effect is mentioned in that sentence. The AE group contains 3,816 sentences, we have 193 sentences for the Dose group, and 6,611 sentences for the NEG group.

Figure 1 represent the distribution of each groups in documents. Figure 1.a shows how many Documents contain either none, 1, 2 or 7 Adverse Effect relations. In continue we randomly select 10% of the documents as Develop data, this part of data is exclusively will be used to do the evaluation while we are training the model Figure 2 shows the distribution on Devel dataset, this part of the data will not be use as part of training. After remove the 10% of Develop we separate another 20% of the data in purpose of isolate from training and evaluation to be used as the test. This will allow us to resemble the situation of using model in real word. Figure 3 shows the distribution of the Test dataset. As we can see we where trying to keep the different situation to let the model have the chance to be tested in all the situation. For example we have some sentences that have 6 adverse drug relation so in devel and test set still we have this situation; however we do not have 7 adverse drug relation because it is so rare to be happen so we keep all of them in training set.

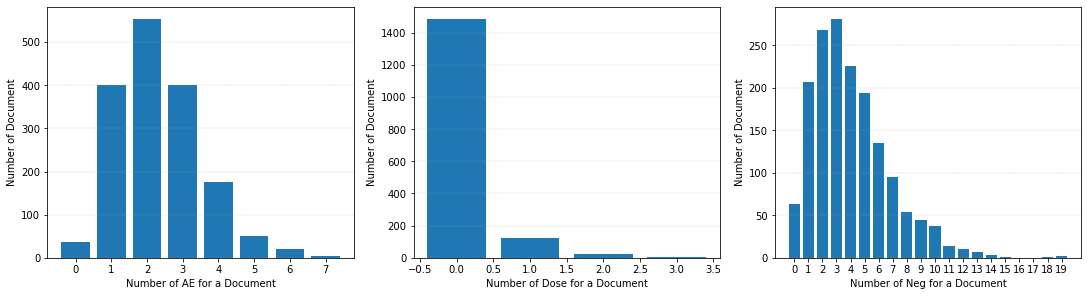


Figure Distribution of on all the Document

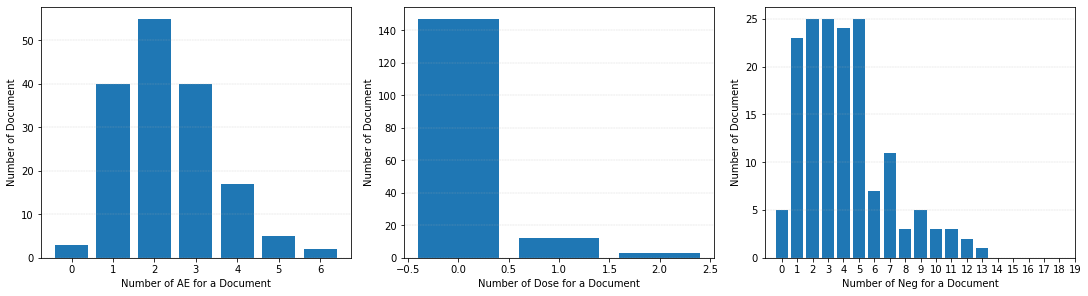


Figure Distribution on Devel

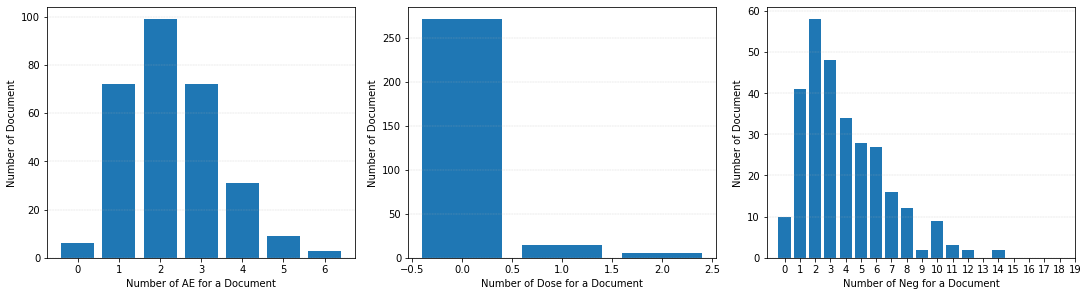


Figure Distribution on Test