Discussão

Conseguimos realizar a extração bem sucedida de uma amostra de documentos de patentes, do qual pré processamos e criamos um corpora que poderá ser usado não somente para este trabalho como para outros trabalhos com documentos de patentes. Fizemos um primeiro levantamento dos tópicos usando o modelo LDA e obtivemos 25 tópicos que serão avaliados e rotulados corretamente.

Utilizando termos simples para rotular os tópicos neste primeiro momento percebemos que dois tópicos puderam ser removidos por serem semelhantes. Com os termos, expandimos com novos termos que são sinônimos, hiperônimos e hipônimos dos termos associados aos tópicos fazendo com que o nosso dicionário cubra uma área maior do assunto.

Por fim, testamos três modelos, muito utilizados na classificação de textos, o Random Forest, Naive Bayes e SVM. O modelo feito em RandomForest obteve um melhor resultado em comparação aos outros dois modelos, onde o modelo final alcançou o valor de acurácia de 0,84, muito mais do que era esperado.

Acreditamos que aplicando outras técnicas de pre processamento e um teste exaustivo de diferentes valores de parâmetros, há a possibilidade de obter um valor de acurácia acima de 0,90.