UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO

Henrique Cursino Vieira

Classificação de patentes utilizando random forest

São Carlos

2020

Henrique Cursino Vieira

Classificação de patentes utilizando random forest

Trabalho de conclusão de curso apresentado ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, como parte dos requisitos para conclusão do MBA em Ciências de Dados.

Área de concentração: Ciências de Dados

Orientador: Prof. Dr. Nikolai Valtchev Kolev

Versão original

São Carlos

2020

"Nenhuma grande descoberta foi feita jamais sem um palpite ousado."

Isaac Newton

LISTA DE ABREVIATURAS E SIGLAS

IDE	Integrated Development Environment
SVM	Support Vector Machine
TF-IDF	Term Frequency - Inverse Document Frequency
FPO	Free Patents Online
DTM	Document-Term Matrix
LDA	Latent Dirichlet allocation

SUMÁRIO

1 INTRODUÇÃO 9
1.1 Apresentação 9
1.2 Justificativa 10
1.3 Problema 10
1.4 Objetivo geral 10
1.5 Objetivos específicos 10
1.6 Metodologia 11

2 DESENVOLVIMENTO 13
2.1 Revisão sistemática 13
2.1.1 Descrição do objeto de estudo 13
2.1.2 Delineamento da pequisa 13
2.2 Materiais e métodos 13
2.2.0.1 Extração dos dados 13
2.2.0.2 Construção do dicionario 13
2.2.1 Validação do dicionario 15
2.2.2 Classificação a partir do dicionario 15
2.2.3 Modelagem 15
2.3 Resultados 15
2.3.1 Extração de dados 15
2.3.2 Construção do dicionario 15
2.3.2.1 Levantamento de tópicos 15
2.3.2.2 Validação dos tópicos 16
2.3.2.3 Expansão do dicionário 16
2.3.2.4 Construção da base de dados 18
2.3.2.5 Construção do modelo 18
2.3.2.5.1 Pre processamento 18
2.4 Discussão 19

3 CONCLUSÃO 21

REFERÊNCIAS 23

1 INTRODUÇÃO

1.1 Apresentação

No desenvolvimento geral de produtos, a pesquisa por documentos de patentes visa garantir o não infringimento de propriedades intelectuais que ainda não estão em domínio público (BREITZMAN; MOGEE, 2002). O sistema de patentes é um conjunto de medidas utilizados para visar o retorno do valor privado investido ao valor social de suas invenções, fornece aos inventores um período temporário de poder de mercado, recuperando os custos de seus investimentos na pesquisa (WILLIAMS, 2017). De acordo com o *World Intellectual Property Indicator 2017*, em 2016, o numero de documentos de patente excedeu 3 milhões pela primeira vez, um aumento de 8.3% (LI, 2018). Em uma pesquisa de documentos de patente, documentos relacionados a tecnologia, economia e jurídico são tratadas, classificadas e analisadas para se obter um alto vantagem técnica e comercial (LI, 2018).

A classificação de documentos é o processo de classificação de um documento em uma categoria predefinida, desempenhando um papel importante no gerenciamento e busca de temas (ANNE et al., 2017). A automatização da classificação de documentos a partir de aprendizado de máquina, pode rotular documentos de um tema único e a rotulagem em vários temas é relativamente desafiador (ANNE et al., 2017).

De acordo com Shahid et al (2019), a classificação de documentos de patente em temas e a atribuição de valor de relevância para estes temas, permitem ao pesquisador filtrar as patentes que o interessa e reduzindo o escopo de analise. Nesse trabalho, realizou a construção de uma matriz de valores de term frequency - inverse document frequency (TF-IDF), notações e peso ponderado por BM25, que posteriormente foi testado em diferentes classificadores, classificando os documentos de patente em cada assunto. Vide Anne et al (2017), identificou uma matriz de métodos a serem aplicados com os modelos k-Nearest Neighbors (kNN), Support Vector Machine (SVM), Random Forest e J48. Os principais passos para essa pesquisa foram técnicas de seleção de características, com uso de ganho de informação e correlação para efetividade do classificadores.

Destes dois estudos, foi observado que a adição de mais características para os modelos de classificação utilizados, a acurácia foi melhorada (SHAHID et al., 2020). E que obstáculos, como o desbalanceamento dos dados foram atenuados pela adição de novas características (ANNE et al., 2017). Balancear a relação entre esses dois pontos é um desafio quanto a classificação de documentos de patente.

1.2 Justificativa

Como tratar, classificar e analisar documentos de patente havendo algumas centenas de documentos sobre um assunto específico? O método tradicional necessita de tempo e equipe para realizá-lo, apresentando um resultado com deficiências devido ao alto volume de documentos de patente a serem analisadas (LI, 2018). Hoje, já há portais web que oferecem ferramentas das quais algumas auxiliam ao pesquisador a reduzir essa pesquisa (ABBAS; ZHANG; KHAN, 2014), mas classificam os documentos em uma relevância geral. Esse resultado somente demonstra que dentro daquela amostra de documentos, uma visão macro sobre o tema que muitas vezes o pesquisador está em busca de um subtema, como quais mercados essa tecnologia está presente, quais os processos de produção desta tecnologia ou qual a formulação desse composto.

1.3 Problema

Com o rápido crescimento de documentos de patente, torna-se urgente a questão de automatização da classificação de documentos de patente de forma acurada e rápida (ZHU et al., 2020). Os documentos de patente contem um potencial conhecimento tecnológico na resolução de problemas no processo de fabricação, nos quais são de grande valor cientifico e tecnológico, no entanto, esse conhecimento está implícito em longos textos (LI, 2018; WANG et al., 2016). A classificação de documentos de patente em temas e subtemas utilizando de modelos de aprendizado de maquina se beneficiaria do uso da extração de características uteis vindas do próprio documento (ANNE et al., 2017). Observa-se que mais de 90% das informações de cientificas e tecnológicas estão em documentos de patente, e sua analise resultaria em decisões de negocio de sucesso (LI, 2018).

1.4 Objetivo geral

Este projeto se propõe a classificar documentos de patente por tema específico e subtemas de interesse do pesquisador, reduzindo o escopo de documentos de patentes a serem estudados à somente os mais relevantes para o que se procura.

1.5 Objetivos específicos

A classificação de documentos de patente envolverá o uso técnicas de processamento de linguagem natural para o tratamento e preparação dos dados que serão usados no modelo de classificação por relevância que será desenvolvido. Este modelo usará inicialmente a medida estatística TF-IDF e avaliaremos outras medidas. Haverá a necessidade de criação de dicionários que auxiliem na classificação dos documentos de patente. E então será treinado um algoritmo para classificar os documentos de acordo com o tema.

1.6 Metodologia

Realizaremos a obtenção de um conjunto de documentos de patente aplicado a agricultura através da ferramenta Free Patents Online - FPO (https://www.freepatentsonline.com/). Não foi encontrado artigos ou materiais que fizessem essa aplicação para patentes relacionadas ao setor agrônomo, para gerenciamento de patentes, desenvolvimento de produtos e descoberta de mercados. Faremos o uso do modelo de classificação baseado em florestas aleatórias, a vantagem desse modelo, é a flexibilidade para o uso em regressão e classificação, além da sua facilidade de interpretação do resultado obtido. A construção de dicionários será a partir de técnicas de Processamento de Linguagem Natural, elencando as palavras mais relacionadas a área. A analise, construção de dicionários e modelagem do modelos de regressão e classificação será feita na linguagem de programação Python.

2 DESENVOLVIMENTO

2.1 Revisão sistemática

O estudo de Revisão Sistemática da Literatura seguiu as recomendações Preferred Reporting Items for Systematic Reviews and Meta-Analisys – PRISMA. Foram buscados os termos: “patent mining”, “patent”, “random forest”, “machine learning” - nas seguintes bases de dados: Periodicos CAPES, Microsoft research, Semantic Scholar e Google Scholar. O intervalo de publicação dos artigos selecionados estão entre 2012 a 2020 e restrito a somente artigos escritos em inglês.

2.1.1 Descrição do objeto de estudo

Foi realizado a extração de dados de documentos de patentes no site Free Patents Online - FPO (https://www.freepatentsonline.com/). Este site contem os dados dos documentos de patentes de forma pública.

2.1.2 Delineamento da pequisa

Foi buscado o termo “agronomy” e filtrado para somente documentos de patentes registrados nos Estados Unidos. Foi totalizado 12906 patentes, dos quais selecionamos uma amostragem das 200 primeiras patentes. Construímos uma aplicação de webscraping na linguagem Python para realizar a extração dos dados de documentos de patentes. Os dados extraídos foram armazenados em um banco de dados.

2.2 Materiais e métodos

2.2.0.1 Extração dos dados

A aplicação de webscraping dos dados de documentos de patentes foi escrita na linguagem de programação Python, com uso das bibliotecas *requests* e *BeautifulSoup*. Essa aplicação é modular o suficiente para que seja definido quantos documentos de patentes terão suas informações extraídas, como também quais informações serão extraídas, como demonstrado na figura 1. Os dados são organizados na forma de tabela e armazenado em um pequeno banco de dados feito em *SQLite*.

2.2.0.2 Construção do dicionario

A construção do dicionario que será utilizado no projeto é composto pelas seguintes etapas, figura 2, geração de um corpora de documentos de patentes, pre processamento do corpora, obtenção da matriz de documento-termo (Document-Term Matrix – DTM) e aplicação do modelo Latent Dirichlet allocation (LDA). A partir dos tópicos apresentados

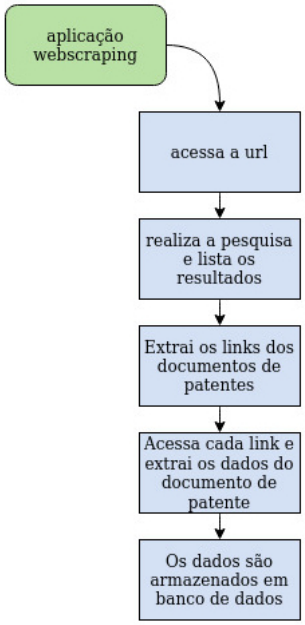

Figura 1: Fluxo de captura dos dados de documentos de patente

pelo resultado do LDA, são adicionados ao tópicos, palavras relacionadas, tais como sinônimos, hiperônimos e hipônimos através do banco de dados wordnet.

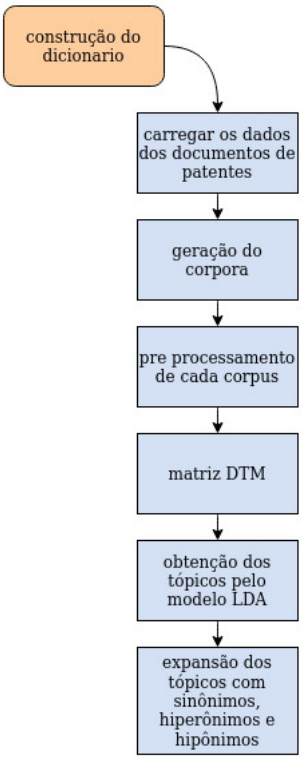

Figura 2: Fluxo de criação do dicionario

2.2.1 Validação do dicionario

A avaliação do dicionario obtido, consiste em observar se o valor k utilizado para geração de tópicos conseguiu separar adequadamente os assuntos contidos no corpora.

2.2.2 Classificação a partir do dicionario

Esta etapa consiste em utilizar o dicionario para classificar os documentos de patentes a partir da iteração com cada termo do dicionario para o conjunto de termos de cada documento, classificando para cada tópico. Esta tarefa é demorada e o tempo necessário aumenta exponencialmente conforme aumenta o tamanho do corpora utilizado. A base de dados gerada será usada para ensinar ao modelo como classificar novos documentos de patentes.

2.2.3 Modelagem

Utilizaremos três dos modelos mais citados na classificação de texto, o RandomForest, Naive Bayes e SVM para avaliar qual se adéqua melhor a essa classificação. Usaremos as técnicas de pré processamento para garantir que o mesmo dado será testado igualmente para cada modelo e escolheremos o modelo de melhor acurácia como modelo final.

2.3 Resultados

2.3.1 Extração de dados

Extraímos uma amostra no total de 904 documentos de patentes através do uso da técnica de webscraping. Destes documentos, os dados de **Titulo** e **Resumo** foram pré processados, removendo as quebras de linhas, espaços no inicio e fim da frase, uso de somente um espaço como separador e transformação do texto em minúsculo. Estes dados foram concatenados e usados para a montagem do corpora de documentos de patentes, que poderá ser utilizado para outros projetos. Podemos visualizar na figura 3 como é distribuida a relação de palavras.

2.3.2 Construção do dicionario

A construção do dicionário engloba o levantamento de tópicos, validação dos tópicos e a expansão do dicionário.

2.3.2.1 Levantamento de tópicos

Foi utilizado o corpora de documentos de patentes feito no passo anterior, onde foi removido as **stopwords** (palavras que não possuem importância a frase, por exemplo em Inglês: The, from, a, an, with, etc.), foram removidos também caracteres numéricos e especiais. O conteúdo de cada corpus foram separados em uma lista palavras, este processo

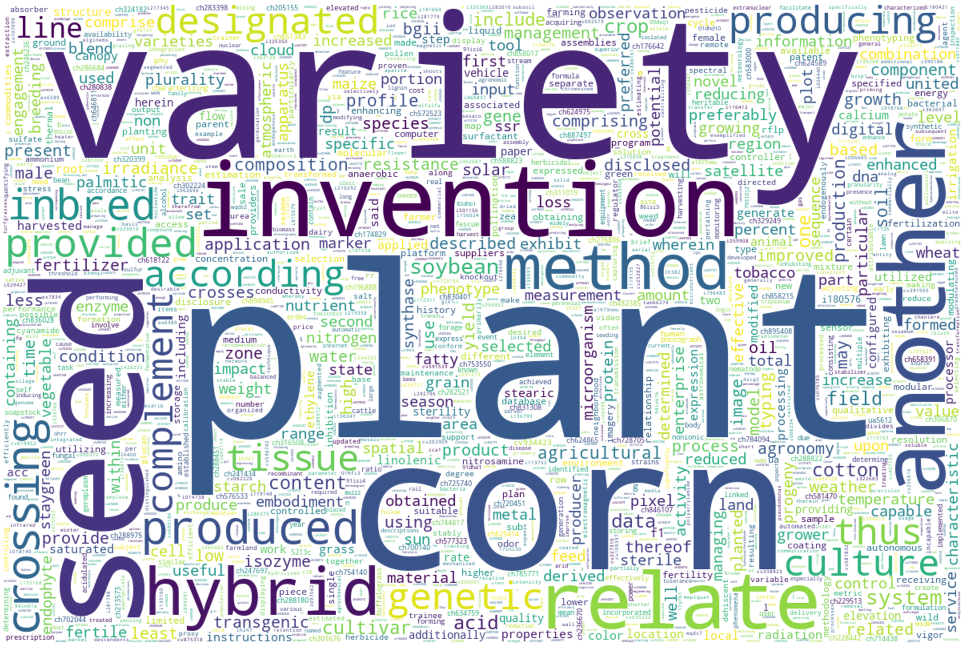

Figura 3: Construção nuvem de palavras dos termos mais representativos para este corpora.

denomina-se como geração de **tokens** e cada token foi desflexionado para a sua palavra raiz (**lemmas**). Obtivemos 904 conjuntos de palavras normalizadas, representando cada documento de patente e que esta pronto para ser utilizado em modelos de Processamento de Linguagem Natural e em modelos de Aprendizado de Maquina. Aplicamos o **modelo LDA**, com os seguintes parâmetros - random_state igual a 100, update_every igual a 1, chuncksize igual a 100, passes igual a 10 e alpha automático. Para definir a quantidade de tópicos k, usamos um laço de 40 interações e anotamos o valor da métrica Coherence.

O gráfico da figura 4 aponta que um k igual a 25 resulta no mais alto valor de Coherence. Utilizaremos este valor para k para se definir os títulos de tópico.

2.3.2.2 Validação dos tópicos

Examinamos o tópicos obtidos através da ferramenta pyLDAvis, figura 5. Os termos que compõe os tópicos gerados representam bem o corpora usado. Temos pouca sobreposição, com exceção do tópico 18, e os termos de cada tópico possuem uma alta relevância com o tema agronomia.

2.3.2.3 Expansão do dicionário

Antes de expandir o dicionário, realizamos a remoção dois tópicos que estavam muito similares. Os tópicos geraram no total de 144 termos únicos que foram submetidas ao wordnet e adicionado os sinônimos, hiperônimos e hipônimos destes termos, totalizando 616 termos que representam cada tópico. A estrutura do dicionário criado é composta por

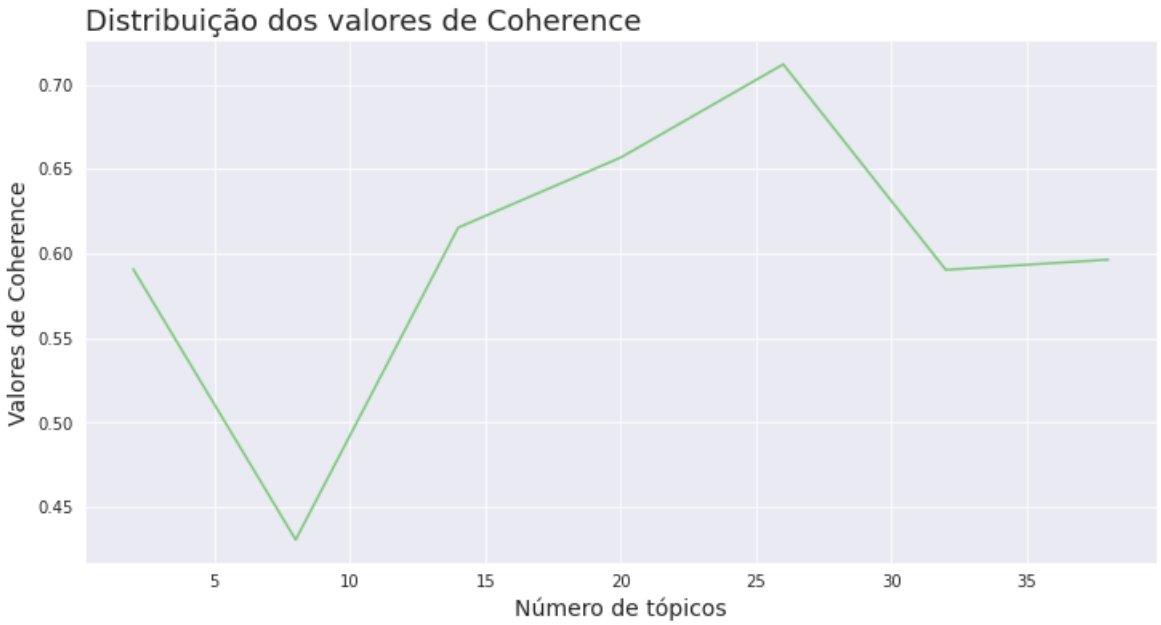

Figura 4: A distribuição dos valores de Coherence ao longo da variação do **parâmetro k**, permite que observemos qual a quantidade de tópicos mais relevantes a se utilizar.

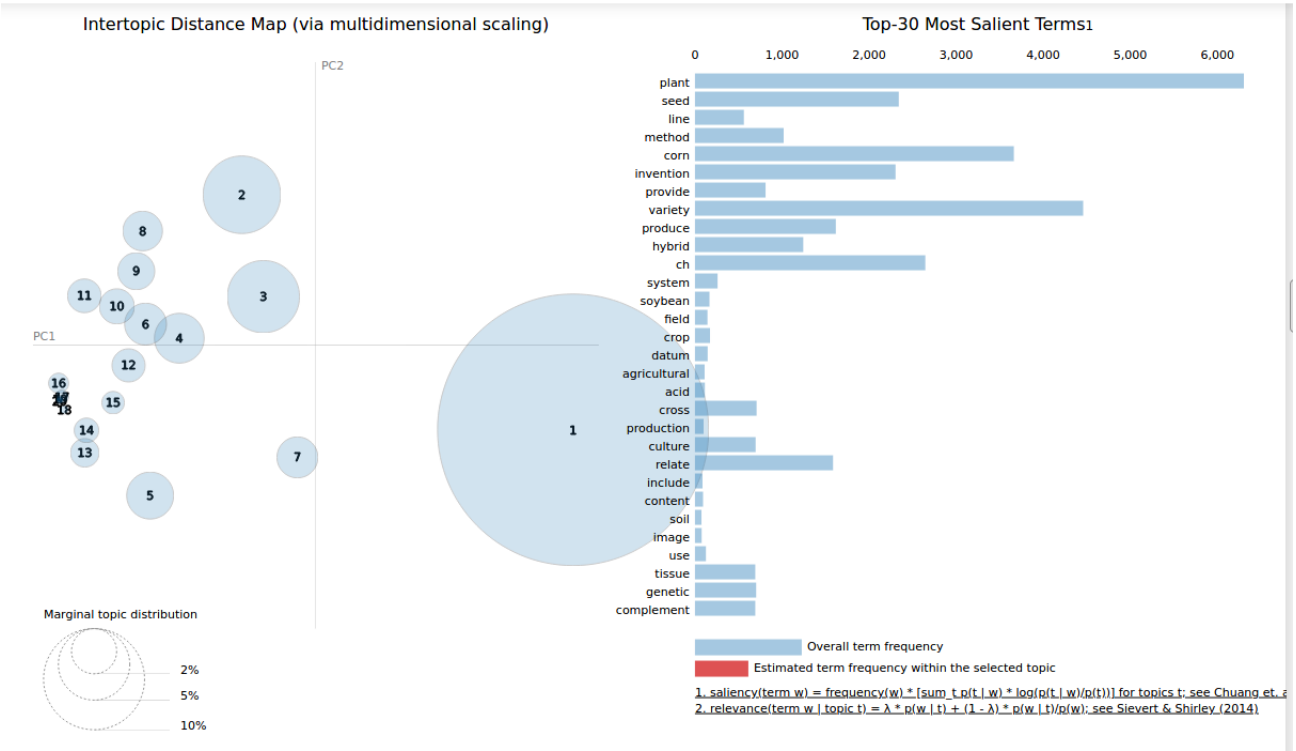

Figura 5: O grafico de bolhas, cada bolha representa um tópico, o tamanho da bolha representa a prevalencia do tópico e a sobreposição de bolhas aponta a similaridade entre os tópicos. O gráfico da direita, as barras representam a relevância do termo para o tópico observado.

três colunas, a primeira é o tópico, a segunda são os termos que estão atrelada ao tópico e a terceira coluna são as palavras derivadas dos termos.

Obtivemos no final um dicionário com com 901 linhas e três colunas, que foi utilizado para fazer uma classificação inicial dos documentos de patentes.

2.3.2.4 Construção da base de dados

A base de dados composta a partir das informações de identificação documento de patente, o título do documento de patente e o seu resumo. A partir do dicionario fizemos uma classificação, onde atribuímos a cada documento de patente os tópicos a que se referem. A estrutura da base de dados é de 817 linhas e 7 colunas, sendo que 87 linhas foram removidas por não conterem título ou resumo.

2.3.2.5 Construção do modelo

Para a construção do modelo principal, os seguintes modelos foram testados, Random Forest, Naive Bayes e SVM, os principais modelos aplicados em classificação de texto. O seguinte fluxo de analise de dados foi aplicado:

2.3.2.5.1 Pre processamento

Matriz documento-termo Conversão da tabela de entrada em uma matriz documento-termo, esta matriz tem a estrutura da seguinte forma:

- colunas: palavras de relevância
- linhas: documentos
- valores: correspondem ao valor de TF-IDF obtido, quando a palavra não consta na entrada, o valor será igual a zero.

A matriz resultante possui 817 linhas e 3492 colunas.

Remoção de stopwords Ao converter a tabela de entrada em uma matriz de documento-termo, as colunas são todas palavras de todos os documentos de patente, fazendo-se necessário a remoção de stop-words, resultando em uma tabela com 817 linhas e 3402 colunas.

Seleção de características A seleção de características tem como objetivo selecionar as colunas que possuam maior relevância a coluna alvo. Utilizamos o método RFE para reduzir o numero de colunas para somente 20. Nossa matriz final possui 817 linhas por 20 colunas.

Modelo Os seguintes parâmetros foram utilizados para cada modelo, seguido do valor de acurácia obtido:

- RandomForest
  - critério de separação: Gini
  - profundidade máxima: 5
  - validação cruzada de 10 folds
  - **acurácia igual a** 0.83
- NaiveBayes
  - foram mantidos o valor padrão para cada parâmetro
  - **acurácia igual a** 0.80
- SVM
  - custo: 1,5
  - **acurácia igual a** 0.78

Avaliação de parametros O modelo escolhido foi o RandomForest por ter o maior valor de acurácia, avaliamos se os parâmetros de profundidade máxima poderia ser melhorado como visto na figura 6.

Após ajustar o parâmetro de máxima profundidade para 15, obtivemos uma acurácia de 0,84. Não houve um ganho significativo em alterar o parâmetro de máxima profundidade.

2.4 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 sobre o tema de agronomia. 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 um valor de acurácia de 0.84, muito mais do que era esperado. Acreditamos que aplicando outras técnicas de pre processamento

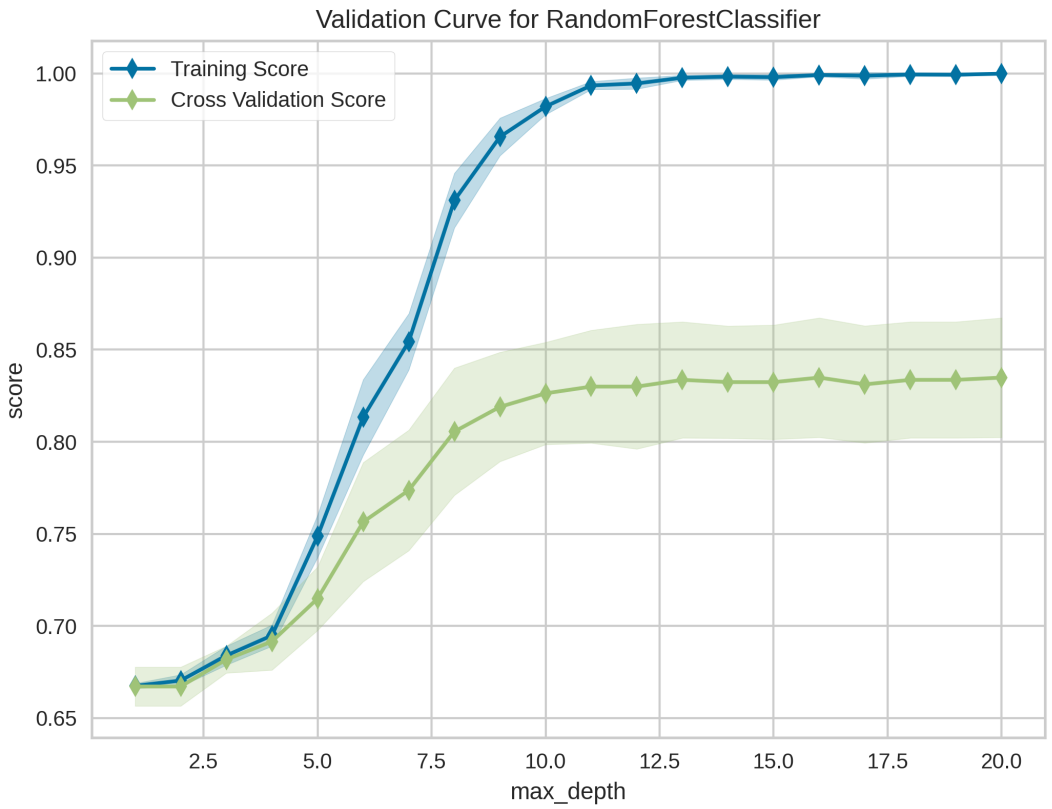

Figura 6: Plot do valor de acurácia obtido para cara valor de profundidade máxima. Podemos observar que para os dados de treino e validação se estabiliza após o valor 10.

e um teste exaustivo de diferentes valores de parâmetros, há a possibilidade de obter um valor de acurácia acima de 0.90.

3 CONCLUSÃO

Ao longo do trabalho demostramos que é possível realizar a classificação dos documentos de patentes a múltiplos tópicos baseado em um dicionario gerado automaticamente a partir do conjunto de documentos de patentes estudado ou que poderá ser baseado em um dicionario construído por um especialista em analise de documentos de patentes. Verificamos que o pre processamento aplicado no tratamento do texto foi satisfatório e o modelo RandomForest teve melhor adequação ao nosso tipo de problema.

Em trabalhos futuros, proponho a melhora da construção de dicionários automaticamente e avaliar se modelos de rede neural teria um melhor resultado de acurácia em relação ao RandomForest.

REFERÊNCIAS

ABBAS, A.; ZHANG, L.; KHAN, S. U. A literature review on the state-of-the-art in patent analysis. **World Patent Information**, Elsevier Ltd, v. 37, p. 3–13, 2014. ISSN 01722190. Disponível em: <http://dx.doi.org/10.1016/j.wpi.2013.12.006>.

ANNE, C. et al. Multiclass patent document classification. **Artificial Intelligence Research**, v. 7, n. 1, p. 1, 2017. ISSN 1927-6974.

BREITZMAN, A. F.; MOGEE, M. E. The many applications of patent analysis. **Journal of Information Science**, v. 28, n. 3, p. 187–205, 2002. ISSN 01655515.

LI, G. A Literature Review on Patent Texts Analysis Techniques. **International Journal of Knowledge and Language Processing**, v. 9, n. 3, p. 1–15, 2018.

SHAHID, M. et al. Automatic patents classification using supervised machine learning. In: SPRINGER. **International Conference on Soft Computing and Data Mining**. [S.l.], 2020. p. 297–307.

WANG, G. et al. Extraction of Principle Knowledge from Process Patents for Manufacturing Process Innovation. **Procedia CIRP**, The Author(s), v. 56, p. 193–198, 2016. ISSN 22128271. Disponível em: <http://dx.doi.org/10.1016/j.procir.2016.10.053>.

WILLIAMS, H. L. How Do Patents Affect Research Investments? **Annual Review of Economics**, v. 9, n. 1, p. 441–469, 2017. ISSN 1941-1383.

ZHU, H. et al. Patent automatic classification based on symmetric hierarchical convolution neural network. **Symmetry**, v. 12, n. 2, p. 1–12, 2020. ISSN 20738994.