Docente: Roberto Hugo Wanderley Pinheiro
Título do Projeto de Pesquisa: Seleção de Características aplicadas a problemas de Classificação de Documentos
Palavras-Chave: Seleção de Características; Aprendizado de Máquina Supervisionado; Classificação de Documentos.

1. INTRODUÇÃO

Este documento apresenta o Projeto de Pesquisa e Plano de Trabalho para o Programa Institucional de Iniciação Científica e Tecnológica (PIICT) do Edital Nº 01/2019/PRPI que será orientado e acompanhado pelo docente Roberto Hugo Wanderley Pinheiro.

1.1.Justificativa

Diversos tipos de problemas de classificação[1] surgem no meio científico sendo que alguns possuem uma coleção vasta de variáveis. Certos problemas precisam lidar com milhares de características[2]. Portanto, muitas dessas aplicações focam seus esforços em diminuir a quantidade de características geradas para tornar a aplicação mais viável. Dentre essas áreas, são costumeiramente citadas: processamento de textos de documentos da Internet, análise de vetores de expressões de gene e química combinatória [1]. Este projeto será focado na seleção de características envolvendo problemas de classificação de documentos.

A classificação de documentos é a tarefa de atribuir a um determinado texto em linguagem natural – em algum idioma específico – uma classe de um universo finito e pré-estabelecido. Aplicações que resolvam esse problema tornam-se cada vez mais difundidas, seja na detecção de spam em e-mails, removendo as mensagens de conteúdo suspeito; organização de documentos em tópicos hierárquicos, para facilitar pesquisas diversas; ou como sistemas de recomendação, indicando certos artigos ou produtos de interesse de um usuário.

Apesar da classificação de documentos ser um problema antigo – estudado desde a década de 1960 – sua importância cresceu apenas na década de 1990, por haver uma grande demanda de aplicações. Com a difusão da Internet, houve um crescimento bastante acelerado na quantidade de informação digital, principalmente devido à facilidade em inserir novos documentos neste meio de comunicação. A necessidade de obter essas informações de modo rápido e prático gerou um interesse da comunidade científica em desenvolver aplicações voltadas para documentos digitais.

[1] Problemas no qual deseja-se atribuir uma classe a exemplos (documentos, pessoas, entidades) ainda não classificados, utilizando o conhecimento obtido por exemplos previamente classificados.
[2] Propriedade individual e mensurável observada nos exemplos. Exemplos: número de patas de um animal, estado civil de uma pessoa, cotação do Euro em determinado dia.

Os problemas de classificação de documentos são considerados diferentes dos problemas tradicionais de aprendizagem de máquina[3] por possuírem duas particularidades desafiadoras: alta dimensionalidade e dados esparsos [2].

A alta dimensionalidade do vetor de característica é dada pela composição do vetor de característica que é, normalmente, de termos (palavras) que aparecem num documento. Deste modo, é comum obter vetores com milhares – ou até mesmo dezenas de milhares – de características, tornando o processo de classificação extremamente custoso e até mesmo inviável em algumas das técnicas, como as redes neurais artificiais.

Enquanto que o problema dos dados esparsos significa que, apesar dos vetores possuírem muitas características, a maioria delas não trazem nenhum peso em determinados documentos, isto é, a matriz composta pelos vetores de características possui poucos campos valorados.

Portanto, a seleção de características é essencial em problemas de classificação de documentos, pois serve para: viabilizar o uso de certas técnicas de classificação, remover as características irrelevantes, incrementar o desempenho da classificação e diminuir a granularidade das características.

1.2.Fundamentação Teórica

Este projeto se divide em duas grandes áreas: seleção de características e classificação de documentos. Estes tópicos serão discutidos nas seções 1.2.1 e 1.2.2, respectivamente.

1.2.1. Seleção de Características

As aplicações com documentos possuem uma grande dimensionalidade, consequentemente grande custo computacional. Uma das soluções para diminuir a dimensão é com o uso da seleção de características. O objetivo desta é selecionar, de um conjunto original T, um conjunto T' de termos que quando usado na classificação dos documentos tenha a maior precisão possível, sendo T' << T [3].

Existem diversas metodologias para realizar uma seleção de características. Os métodos podem ser divididos em dois tipos: métodos de filtragem [4] e métodos *wrappers* [5].

Os métodos *wrappers* utilizam classificadores para testar diversos subconjuntos de T, visando obter o melhor subconjunto de características baseado na performance[4] obtida pelo classificador escolhido. Apesar de garantirem uma boa seleção, sua aplicação é impraticável em problemas com grandes dimensões devido ao elevado custo computacional exigido pelo uso constante de um classificador na avaliação de cada subconjunto.

Por outro lado, os métodos de filtragem são mais velozes que os métodos *wrappers*, pois esta abordagem não requer a presença do classificador para avaliar o subconjunto de características selecionadas. Apesar de não atingirem os mesmos resultados dos métodos

[3] Campo da ciência da computação que possibilita o computador aprender sem ser programado especificamente para tal.
[4] Avaliação do quão bem o classificador cumpre a tarefa de classificação.

wrappers, os métodos de filtragem são mais apropriados para problemas de classificação de documentos, por conseguirem realizar uma seleção de modo rápido e eficiente.

Dentre os métodos de filtragem utilizados na seleção de características, para problemas de classificação de documentos, o algoritmo mais utilizado é o *Variable Ranking* [1] por conta de seus resultados consistentes. O funcionamento do *Variable Ranking* se resume em selecionar n (parâmetro de entrada) características que sejam consideradas as melhores, com base numa função de avaliação de características [6] [7]. Estas funções medem a importância de cada característica existente no problema, dentro daquela base de dados estudada. Isto é, medem a capacidade de determinada característica em oferecer informações suficientes para que o classificador consiga classificar corretamente duas ou mais classes.

Diversas pesquisas permanecem sendo realizadas para criação de novos algoritmos de seleção de características mais eficientes [8] [9] [10]. Apesar destes novos métodos conseguirem incrementar os resultados, o *Variable Ranking* permanece popular devido à constante criação de novas funções de avaliação de características [7] [11], que melhoram sua performance cada vez mais. Adicionalmente, o tempo de processamento do *Variable Ranking* é inferior aos métodos mais complexos, pois requer apenas um cálculo para cada característica e uma ordenação delas, de acordo com a função de avaliação de características.

Em 2012, um estudo foi feito para criar um algoritmo tão simples quanto o *Variable Ranking* e que ainda assim conseguisse aprimorar seus resultados [12]. Para tal, foi observado que o algoritmo *Variable Ranking* tem uma dificuldade: encontrar o melhor valor para n não é uma tarefa trivial. Neste caso, é necessário testar várias possibilidades para n, transformando o algoritmo num método *wrapper*. Deste modo, o algoritmo perde sua principal vantagem que é o rápido processamento. Uma segunda opção é utilizar valores já encontrados por outros trabalhos, mas esta opção não isenta o algoritmo de sua falha, apenas facilita o trabalho do pesquisador.

Baseado na desvantagem do algoritmo *Variable Ranking*, foi criado o *At Least One FeaTure* (ALOFT) [12], método que seleciona as características de modo automático – sem a necessidade do parâmetro *n* – e utiliza as funções de avaliação de características, garantindo sua evolução e aplicabilidade junto com novas funções que venham a ser criadas.

1.2.2. Classificação de Documentos

Segundo [3] a tarefa de classificação de documentos se resume a designar um valor booleano ao par $\langle d_j, c_i \rangle \in \mathcal{D} \times \mathcal{C}$, no qual $\mathcal{D}$ é conjunto de todos os documentos e $\mathcal{C} = \{c_1, \dots, c_N\}$ é um conjunto pré-definido de N classes. Um valor *true* em $\langle d_j, c_i \rangle$ indica que o documento dj pertence a classe c_i, enquanto um valor *false* indica que o documento d_j não pertence a classe c_i. Uma classificação pode ser do domínio binário ou multi-classe. Esses domínios representam restrições estabelecidas pela classificação final do problema.

A classificação binária é o caso fundamental. Domínios binários são comuns e generalistas, pois problemas multi-classe podem ser resolvidos dividindo-os em diversos

problemas binários. Sua tarefa é simples: classificar um determinado documento em uma das duas classes existentes, normalmente chamadas classe positiva e classe negativa.

Com relação aos domínios multi-classe, existem duas maneiras distintas de classificação. A primeira é conhecida como classificação *single-label* (etiquetagem-única), no qual o documento é classificado como pertencente, obrigatoriamente, a uma das classes existentes, sendo uma classificação de um para n. A segunda forma é conhecida como classificação *multi-label* (etiquetagem-múltipla) que visa classificar um documento como pertencente a uma, nenhuma, várias ou até mesmo todas as classes existentes.

Os classificadores mais utilizados para classificação de documentos são o *k-Nearest Neighbor* (kNN), Naive Bayes e o *Support Vector Machine* (SVM).

O kNN é bastante utilizado devido à sua facilidade de implementação e por ser afetado drasticamente pela seleção de características, tanto no seu tempo computacional como na qualidade da classificação. Entretanto, a implementação utilizada é modificada para um melhor aproveitamento [13]. A versão modificada do kNN, para classificação de documentos, utiliza a medida de similaridade do cosseno – para encontrar os vizinhos – em vez da clássica distância euclidiana.

Assim como o kNN, o Naive Bayes também é um classificador de simples implementação e possui duas versões de uso específico para problemas de classificação de documentos: *Multi-Variate Bernoulli Event Model* e *Multinomial Event Model* [14]. O primeiro modelo utiliza apenas a característica de ausência ou presença do termo no documento (valor binário), enquanto que a segunda utiliza a frequência do termo no documento (valor inteiro). Ambos os modelos modificam a maneira como as probabilidades são calculadas utilizando especificidades do problema como tamanho do vocabulário e aparições das palavras.

Por fim, o SVM é bastante utilizado em classificação de documentos [13] [16], sendo considerado o classificador estado-da-arte para o problema em estudo. Entretanto, diferentemente do kNN e Naive Bayes, o SVM consegue lidar bem com problemas de alta dimensionalidade [17]. Portanto, o SVM raramente é utilizado para verificar a qualidade de uma seleção de características, sendo este o papel desempenhado pelo kNN e Naive Bayes, cujo impacto causado pela seleção é bastante perceptível [11] [13].

2. OBJETIVOS

2.1. Objetivo Geral

Este trabalho tem por objetivo principal propor, no mínimo, uma nova técnica ou abordagem para selecionar características, e avaliar seus resultados pela comparação com técnicas estado-da-arte.

2.2. Objetivos Específicos

- Revisão aprofundada da literatura para focalizar os conhecimentos neste problema específico;

- Estudar algoritmos de Aprendizagem de Máquina visando o conhecimento de diversos classificadores para validar a seleção realizada futuramente;
- Implementar os classificadores;
- Estudar técnicas de seleção de características;
- Propor modificações ou novas abordagens para o problema em questão, seja na parte da classificação ou na seleção de características;
- Comparar o desempenho das técnicas clássicas com as propostas;
- Analisar resultados obtidos.

3. METODOLOGIA

Neste projeto, serão conduzidos estudos de modo a incentivar o desenvolvimento de um método de seleção de características. Deste modo, os trabalhos serão iniciados com revisão da literatura de Classificação de Documentos [2] [3] e Seleção de Características [4] [6]. A revisão da literatura servirá para aprendizagem do conhecimento da área de estudo e para construir uma base necessária para dar procedimento ao projeto. A revisão será realizada principalmente pela bibliografia disponível a partir da CAPES, em buscas com palavras-chaves como *feature selection*, *text categorization* e *filtering methods*. Uma vez concluída a revisão da literatura, os alunos envolvidos nesse projeto deverão solidificar esse conhecimento, formalizando e implementando alguns dos métodos clássicos encontrados na literatura [1] [12]. Desta forma, os alunos poderão amadurecer o conhecimento prático e teórico na área para, finalmente, iniciarem suas contribuições na área.

Todo esse processo será acompanhado com reuniões presenciais dos bolsistas e do orientador (docente) com exposição do progresso do trabalho e apresentações para aprendizado mútuo. Essas atividades deverão respeitar um cronograma de execução do projeto:

Atividade	**Mês**											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Revisar Literatura	X	X	X									
Documentar Revisão da Literatura	X	X	X									
Estudo Compartilhado			X	X								
Implementar Literatura				X	X	X	X					
Elaborar Proposta						X	X					
Implementar Proposta								X	X	X		
Executar Experimentos									X	X	X	

Analisar Resultados										X	X	
Escrever Artigo										X	X	X
Relatório Final										X	X	X

Atividades do Cronograma são descritas a seguir:

- Revisar Literatura: Um estudo do estado-da-arte no domínio da seleção de características e das técnicas de classificação para aprofundar o conhecimento na área e conhecer os trabalhos mais recentes.
- Documentar Revisão da Literatura: Documentar as técnicas e metodologias estudadas, guardar referências importantes para futuros trabalhos, documentar ideias de novas propostas ou planos de propostas. Essa documentação será utilizada para elaborar os relatórios de atividades que deverão ser apresentados ao final de cada ano.
- Estudo Compartilhado: com o assunto mais amadurecido por ambos os bolsistas, estes deverão compartilhar o material estudado, pois o projeto requer integração de ambos os bolsistas e compreensão completa de todo escopo do projeto.
- Implementar Literatura: Implementar as técnicas e métodos estudados na revisão da literatura. Cada bolsista será responsável por uma parte do projeto, sendo ambas implementações partes de um todo e compondo todo um Sistema de Classificação de Documentos.
- Elaborar Proposta: Investigar novas maneiras de melhorar os resultados das técnicas clássicas para criar novas técnicas ou elaborar melhorias significantes, nas já existentes.
- Implementar Propostas: Implementar as ideias desenvolvidas na elaboração de propostas.
- Executar Experimentos: Selecionar diversas bases de dados consolidadas no domínio de aplicação para comparar os resultados obtidos nas implementações da literatura.
- Analisar Resultados: Fazer uma análise aprofundada dos resultados obtidos.
- Escrever Artigo: Condensar as novas propostas num artigo científico. O tempo necessário para conclusão deste artigo, será utilizado para sua elaboração, revisão e submissão.
- Relatório Final: Preparar um documento baseado em toda a documentação de estudo elaborada até o momento, e com os resultados e conclusões relevantes do artigo. Deste modo, todo o conteúdo produzido e estudado será reunido de modo coerente e contínuo, demonstrando o conhecimento adquirido na área e as contribuições para o domínio estudado.

As implementações serão desenvolvidas em Python 3.x, facilitados pelo pacote *Anaconda*[5], ambiente de desenvolvimento *Spyder* e biblioteca de Aprendizagem de Máquina

[5] https://anaconda.org/anaconda/python

scikit-learn[6]. Testes estatísticos para avaliação das propostas e comparações com outros métodos clássicos da literatura serão realizados com o programa estatístico R[7].

Para atingir os objetivos deste projeto, será necessário a participação de bolsistas remunerado orientados em Iniciação Científica (PIBIC), nomeados bolsista A, focado na Seleção de Características e bolsista B, focado na Classificação de Documentos. O trabalho em equipe de ambos os bolsistas possibilitará a execução deste projeto com acompanhamento do docente para que as tarefas sejam cumpridas nos prazos pré-estabelecidos. Para atingir os objetivos deste projeto, seguiremos uma sequência de atividades com prazos estabelecidos de acordo com um cronograma. Segue descrição das atividades gerais:

4. CONTRIBUIÇÕES CIENTÍFICAS

Espera-se, com a execução completa deste projeto, a elaboração de um Sistema de Classificação de Documentos completo em Python, com todas as etapas necessárias para seu funcionamento. Adicionalmente, os métodos propostos durante este projeto serão utilizados para escrita de artigos para publicação em períodos de ampla circulação e eventos da área. Todo o material produzido beneficiará futuras pesquisas.

5. REFERÊNCIAS BIBLIOGRÁFICAS

[1] I. Guyon & A. Elisseeff. An introduction to variable and feature selection. Journal of Machine Learning Research. 1157–1182. 2003.

[2] J. Su, J. Sayyad-Shirabad, S. Matwin. Large Scale Text Classification using Semi-supervised Multinomial Naive Bayes. Proceedings of the 28th Internacional Conference on Machine Learning. 2011.

[3] F. Sebastiani. Machine learning in automated text categorization. ACM Comput. Surveys. 34(1):1–47. 2002.

[4] G.H. John, R. Kohavi, K. Pfleger. Irrelevant features and the subset selection problem. Proceedings of the 11th International Conference on Machine Learning. 121–129. 1994.

[5] I. Moulinier, & J. Ganascia. Applying an existing machine learning algorithm to text categorization. Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing. 343–354. 1996

[6] Y. Yang & J. Pedersen. A comparative study on feature selection in text categorization. Internacional conference on machine learning. Proceedings of the 14th International Conference on Machine Learning. 412–420. 1997.

[6] http://scikit-learn.org/
[7] https://www.r-project.org/

[7] G. Forman. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. Journal of Machine Learning Research, 3:1289–1305. 2003.

[8] M. Rogati & Y. Yang. High-performing feature selection for text classification. Proceedings of the 11th International Conference on Information and Knowledge Management. 659–661. 2002.

[9] W. Shang, H. Huang, H. Zhu, Y. Lin, Y. Qu, Z. Wang. A novel feature selection algorithm for text categorization. Expert Systems with Applications. 2007.

[10] J. Yang, Y. Liu, X. Zhu, Z. Liu, X. Zhang. A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization. Information Processing & Management. 741–754. 2012.

[11] J. Chen, H. Huang, S. Tian, Y. Qu. Feature selection for text classifcation with Naive Bayes. Expert Systems with Applications. 5432–5435. 2009.

[12] R.H.W. Pinheiro, G.D.C. Cavalcanti, R. Correa & I.R. Tsang. A global-ranking local feature selectiom method for text classification. Expert Systems with Applications. 12851–12857. 2012.

[13] X. Xue & Z. Zhou. Distributional Features for Text Categorization. IEEE Transactions on Knowledge and Data Engineering. 21, 3, 428–442. 2009

[14] A. McCallum & K. Nigam. A comparison of event models for naive bayes text classification. Workshop on learning for text categorization. 41-48. 1998.

[16] T. Joachims. Text categorization with support vector machines: learning with many relevant features. Proceedings of 10th European Conference on Machine Learning. 137–142. 1998.

[17] Y. Yang & X. Liu. A Re-Examination of Text Categorization Methods. Proceedings of 22th Conference on Research and Development in Information Retrieval. 42–49. 1999.