**Desafio 01 – Resumo de Artigo Aplicado a Processamento de Linguagem Natural**

Para a entrega do primeiro desafio proposto na aula de Processamento de Linguagem Natural, foi escolhida a tese de mestrado *“Estudo Comparativo de Análise de Sentimentos Aplicado à Notícias Políticas”*, desenvolvido por Caio Magno Aguiar de Carvalho em Fevereiro de 2018 sob orientação de Allan Kardec Barros e Ewaldo Santana pela Universidade Federal do Maranhão.

A motivação dessa escolha foi dada pela presença de elementos extremamente interessantes dentro do âmbito de Processamento de Linguagem Natural, como a classificação e a análise de sentimentos, bem como a utilização de técnicas específicas de seleção de atributos afim de melhorar a performance de modelos supervisionados de classificação de texto. O contexto utilizado diz respeito a um *Corpus* coletado com notícias de cunho político veiculadas por grandes meios de comunicação, tornando este um trabalho extremamente atrativo e que, de fato, despertou meu interesse.

No trabalho, é proposto um estudo comparativo entre as técnicas de representação de texto, seleção de atributos e ferramentas de aprendizagem de máquina para se classificar notícias políticas coletadas em portais online sobre as eleições brasileiras de 2014 quanto a sua opinião/sentimento (positivo, negativo ou neutro). Neste estudo, os classificadores Naive Bayes, Support Vector Machines e Regressão Logística foram avaliados juntamente com as técnicas de seleção de atributos Qui Quadrado, Categorical Proportional Difference e Categorical Probability Proportional Difference.

Com relação ao fluxo de desenvolvimento e organização da tese, tem-se um início citando as principais motivações do autor para escolha do tema, sempre relacionando a opinião pública como uma grande ferramenta para tomadas de decisões, sejam essas vinculadas aos campos de vendas online, críticas cinematográficas e política. Neste último, as acusações de imparcialidade dos grandes meios de comunicação pode ser considerada como uma grande inspiração para a realização do trabalho, visto que alegações de manipulação da opinião pública realmente necessitam ser comprovadas e isentas de qualquer viés. Assim, técnicas de PLN foram aplicadas dentro deste contexto político para colaborar com pesquisas de classificação de notícias com o objetivo de comparar o desempenho de classificadores a partir de diferentes métodos de seleção de atributos.

É interessante citar a sessão de “Trabalhos Relacionados” da teste, dado que alguns trabalhos importantes como Morgado (2012), Martinazzo, Dosciatti e Paraiso (2011) e Alvim et al. (2010) foram utilizados como exemplos. Cada um deles aplicou diferentes técnicas em diferentes contextos para classificação de notícias, como a utilização de unigramas, bigramas, POS tagging, SVM, LSA, entre outros.

Na sequência, o autor descreve uma completa Fundamentação Teórica de todas as técnicas e modelos utilizados ao longo do trabalho. As explicações nessa sessão descrevem os seguintes conceitos: bag-of-words, modelos de aprendizagem supervisionada (Naive Bayes, Regressão Logística e SVM) e, por fim, métodos de seleção de atributos (Teste Qui Quadrado, Diferença Proporcional Categórica e Diferença Proporcional Categórica Probabilística).

Com a leitura da teste, foi possível aprender novos conceitos relacionados a seleção de atributos. O Teste Qui Quadrado mede o nível de dependência estatística entre duas variáveis aleatórias: a distribuição de um determinado ngram ao logo dos textos e as suas respectivas classes. O valor qui quadrado assume o valor 0 quando as duas variáveis são estatisticamente independentes e cresce a medida que as variáveis tornam dependentes uma da outra. Esse valor é utilizado para selecionar os ngrams que estão mais relacionados com uma determinada classe. Quanto maior o valor qui quadrado mais relevante é esse ngram para aquela classe.

Já o método da Diferença Proporcional Categórica, ou CPD, é uma técnica supervisionada que atribui para cada palavra um peso de importância para cada categoria de interesse da classificação. Nesse método, defende-se que a frequência de palavras específicas de um certo assunto caracteriza o tópico majoritário daquele texto. O valor CPD de uma palavra para uma categoria mede o quanto uma palavra contribui para diferenciar uma categoria das outras dentro do mesmo corpus. Se ela ocorre muito em documentos de uma determinada categoria, ela é um forte indicador desta classe. Se ela ocorre de modo igualmente disperso entre as categorias de textos, essa palavra não carrega muita informação para distinção das categorias existentes. O valor calculado é compreendido no intervalo [-1, 1], onde 1 indica que aquela palavra ocorre somente em documentos daquela categoria e -1 indica que a aquela palavra ocorre em todas as outras categorias menos naquela categoria de interesse. O valor CPD de uma palavra é o valor CPD da palavra para a categoria cujo valor é máximo.

Entretanto, é possível afirmar que o valor CPD não é uma maneira ótima de considerar a importância de ngrams para cada classe. Dito isso, o método da Diferença Proporcional Categórica Probabilística (CPPD) é proposto como uma forma de melhoria do CPD. Nele, além de calcular o valor CPD de cada ngram, o CPPD ordena os ngrams por sua respectiva probabilidade de ocorrência dentro da classe.

Após as exemplificações teóricas, o autor seguiu com a implementação prática do projeto e, utilizando a linguagem Python, trabalhou em uma base de dados gerada em formato XML (inserida posteriormente em um banco de dados MySQL) para classificar notícias entre classes negativa, positiva ou neutra a partir de diferentes métodos de seleção de atributo. As notícias foram coletadas pelo twitter e englobam os meios de comunicação Veja, Estadão, Folha, G1 e Carta Capital. Como preparação da base, apenas foram removidas as *stopwords*, convertidos os tokens para minúsculo e substituídos números pelo token “NUMBER”.

Dessa forma, o autor aplicou o treinamento e mediu a acurácia dos modelos Naive Bayes, Regressão Logística e SVM sem a aplicação de métodos de seleção de atributos e comparou este resultado com a aplicação dos métodos de seleção de atributo mencionados acima. Como resultado, foi obtido um melhor desempenho com o modelo de Regressão Logística, utilizando *unibigrams* que, apesar de proporcionar um aumento na dimensionalidade do problema, realmente apresentou melhores resultados, e a técnica de seleção de atributos CPPD. Comparando a acurácia média obtida, o modelo de Regressão Logística sem a aplicação de métodos de seleção de atributo apresentou uma acurácia de 57%. Já após a aplicação do método CPPD, essa acurácia subiu para 84,93%