CATEGORIZAÇÃO AUTOMÁTICA DE DOCUMENTOS UTILIZANDO PROCESSAMENTO DE LINGUAGEM NATURAL E ALGORITMOS DE AGRUPAMENTO

Lucas Vilas Boas Alves[1]; Erick Giovani Sperandio Nascimento[2]
[1] Bolsista; Pesquisa, Desenvolvimento e Inovação (PD&I); Projeto Centro de Competência em Inteligência Artificial; lucas.alves@fbter.org.br
[2] Doutorado; Centro Universitário SENAI CIMATEC; Salvador-BA; erick.sperandio@fieb.org.br

RESUMO

O objetivo deste trabalho é avaliar o uso de técnicas de processamento de linguagem natural associadas a algoritmos de agrupamento para categorizar documentos de texto de forma automática. O estudo foi baseado em um conjunto de dados com mais de 45.000 documentos contendo notícias de um jornal do Espírito Santo. As categorias atribuídas pelo jornal às notícias foram usadas como referência para mensurar a qualidade dos *clusters* produzidos, dado que notícias de uma mesma categoria, por serem similares, devem ser aglomeradas em um mesmo grupo. A metodologia proposta foi avaliada usando estatísticas como acurácia, precisão, sensibilidade e F1-score, comumente utilizadas para mensurar a qualidade de modelos de classificação.

PALAVRAS-CHAVE: categorização automática; documentos textuais; processamento de linguagem natural

1. INTRODUÇÃO

Todos os anos são produzidos grandes volumes de novos documentos de texto digital, que se acumulam aos já disponibilizados em meios como a internet, bibliotecas digitais e fontes de notícias. Este crescimento constante acarreta na necessidade de desenvolver métodos que possam auxiliar os usuários a navegar, resumir e organizar estas informações com o objetivo final de ajudá-los a encontrar o que estão procurando. Os algoritmos de agrupamento de documentos desempenham um papel importante em relação a esse objetivo, pois fornecem um mecanismo de navegação e pesquisa, organizando grandes quantidades de informações em *clusters* significativos.[1] Estas técnicas são comumente aplicadas na organização automática de documentos, extração de tópicos e recuperação de informações, tendo como objetivo reunir documentos semelhantes em categorias, onde a semelhança é determinada a partir de alguma métrica que leva em conta as características de um documento.

Dados de linguagem natural humana, como a fala ou textos, não são estruturados, isto é, não seguem uma estrutura regular rígida e previamente definida que facilite a recuperação de informações. Isto torna necessário a aplicação de técnicas de processamento de linguagem natural para adequá-los a um formato conveniente para a computação e obtenção das informações desejadas. Quando se trata de textos, um dos recursos mais comuns é transformar os documentos em uma representação numérica, na forma de um vetor. Estes vetores numéricos podem em seguida ser aplicados a algoritmos de aprendizado de máquina para, por exemplo, classificá-los em categorias previamente definidas.

O presente trabalho tem como propósito avaliar o uso de técnicas de processamento de linguagem natural associadas com algoritmos de agrupamento para categorizar documentos de texto. O estudo foi baseado em um conjunto de dados com mais de 45.000 documentos contendo notícias de um jornal do Espírito Santo. As categorias atribuídas pelo jornal às notícias foram usadas como referência para mensurar a qualidade dos *clusters* produzidos, dado que notícias de uma mesma categoria, por serem similares, devem ser aglomeradas em um mesmo grupo.

2. METODOLOGIA

Os documentos foram inicialmente pré-processados para retirar tags HTML e separar metadados contidos no corpo dos textos, como a data de publicação e a categoria da notícia. Também foram identificados e eliminados do conjunto de dados as amostras cujos textos eram repetições, mantendo apenas documentos com textos distintos. Em seguida, as publicações passaram por um processo de tokenização, onde os espaços em branco dos textos foram utilizados para separar elementos textuais chamados de tokens. Cada um destes elementos passou posteriormente por um processo de tratamento de letras maiúsculas, pontuações, símbolos e números contidos nos textos. Por fim, foram filtrados tokens muito frequentes, tais como preposições, artigos ou conjunções, e tokens pouco frequentes, como palavras com a grafia errada, nomes próprios, modelos de equipamentos ou outras palavras muito específicas. Estes tokens, em geral, carregam pouco significado e não trazem características relevantes que auxiliem no processo de classificação.[2]

Na etapa seguinte, os documentos, agora caracterizados por uma sequência de tokens, foram transformados em vetores numéricos utilizando uma representação conhecida na literatura como *Bag of Words*. Neste modelo a ordem exata dos termos de um texto é ignorada e cada documento se torna um vetor com um valor para cada termo do vocabulário. Os pesos do vetor de cada documento foram então calculados utilizando a técnica TF-IDF, abreviação de *Term Frequency - Inverse Document Frequency*, que se destina a refletir a importância de uma palavra dentro de um documento levando em conta uma coleção de documentos. Este método atribui um peso maior ao termo que ocorre várias vezes em um pequeno número de documentos, e um peso menor ao termo que ocorre poucas vezes em um documento ou que ocorre em muitos documentos.[3]

Com os textos representados na forma de vetores numéricos, foi possível então aplicar o algoritmo de agrupamento hierárquico aglomerativo para produzir os *clusters* de documentos. Este algoritmo usa uma abordagem ascendente em que cada amostra começa em seu próprio *cluster* e, em seguida, os *clusters* são mesclados sucessivamente com base no critério de ligação, que determina a métrica de similaridade usada na estratégia de mesclagem. Essa hierarquia de *clusters* é representada como uma árvore, onde a raiz da árvore é o agrupamento único que reúne todas as amostras e as folhas são os agrupamentos com apenas uma amostra.[4] Como os documentos estavam divididos em 21 categorias jornalísticas, o algoritmo de agrupamento foi configurado para produzir o mesmo número de *clusters*.

Diversos experimentos foram realizados com o intuito de avaliar como a variação de alguns parâmetros impactaria no desempenho da categorização. Uma das variáveis testadas foi a alteração do tamanho do vocabulário, utilizado na construção dos vetores dos documentos, a partir da eliminação maior ou menor de termos pouco frequentes. Outras duas variáveis investigadas foram o critério de ligação e a métrica de similaridade, ambas relacionadas ao algoritmo de agrupamento hierárquico. Finalmente, também foi experimentada a aplicação da técnica *Singular Value Decomposition* (SVD), com a variação do número de componentes principais a serem mantidos, tendo como objetivo reduzir a dimensionalidade dos vetores dos documentos.

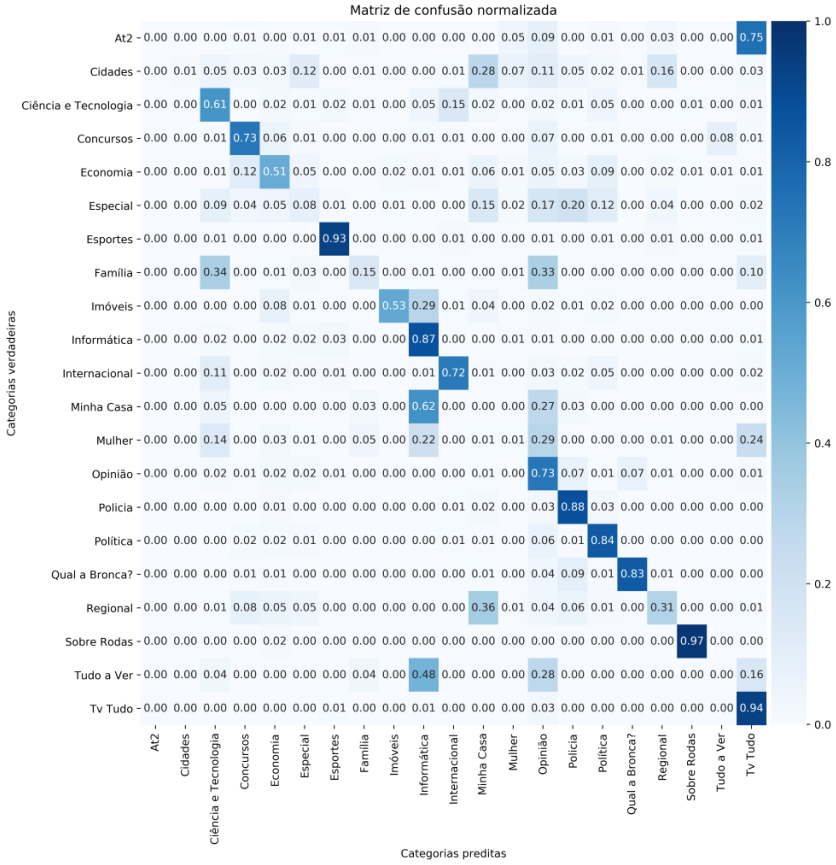
3. RESULTADOS E DISCUSSÃO

Com os testes foi possível observar que, até certo ponto, a diminuição do tamanho do vocabulário, através da eliminação de termos pouco frequentes, produziu uma melhor categorização dos textos. Os resultados também foram melhores quando a dimensionalidade dos vetores foi reduzida, através do uso da técnica de SVD. Não foi possível perceber nenhum padrão claro relacionado aos critérios de ligação e as métricas de similaridade, dado que as melhores combinações variaram dependendo de características na representação dos vetores.

Para calcular as métricas comumente utilizadas em problemas de classificação, cada uma das 21 categorias do jornal foi associada a um dos 21 *clusters* encontrados pelo algoritmo de agrupamento. O experimento com os melhores resultados utilizou vetores produzidos com um vocabulário de 21.590 termos e reduzidos com SVD para 20 componentes. O critério de ligação utilizado no algoritmo de agrupamento hierárquico foi a média das distâncias (*average linkage*) e a métrica de similaridade aplicada foi a distância cosseno. Como resultado atingiu-se uma acurácia de 56%, precisão média ponderada de 60%, sensibilidade média ponderada de 56% e F1-score médio ponderado de 52%.

A Figura 1 exibe a matriz de confusão normalizada, que demonstra como os textos foram categorizados. O bom desempenho atingido na classificação das categorias “Esportes” e “Sobre Rodas” pode estar ligado ao fato dessas categorias terem conteúdos mais específicos, enquanto categorias como “At2” e “Tv Tudo” podem ter obtido um desempenho pior por tratarem de conteúdos semelhantes.

Figura 1 – Matriz de confusão normalizada pelo número de elementos de cada classe



4. CONSIDERAÇÕES FINAIS

As métricas encontradas demonstram que a metodologia proposta, baseada em técnicas simples e consolidadas, foi capaz de obter resultados satisfatórios na categorização dos documentos. Um ponto a ser observado é que algumas categorias do jornal agrupam notícias com conteúdo genérico e muitas vezes similar à de outras categorias. Isto pode explicar as características de alguns clusters produzidos pelo algoritmo e uma possível razão para melhores resultados não terem sido alcançados com os métodos aplicados.

Agradecimentos
Este trabalho contou com o apoio da FAPESB (Fundação de Amparo à Pesquisa do Estado da Bahia), na forma de uma bolsa de Desenvolvimento de Inovação Tecnológica no projeto "Implantação de Infraestrutura de Pesquisa em Simulação e Modelagem Computacional no Estado da Bahia Utilizando Processamento de Alto Desempenho", desenvolvido no Centro de Supercomputação para Inovação Industrial do SENAI CIMATEC.

5. REFERÊNCIAS

[1] ZHAO, Ying, KARYPIS, George. **Criterion Functions for Document Clustering:** Experiments and Analysis. 2002.
[2] SAIF, Hassan et al. **On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter.** In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pp. 810-817, 2014.
[3] MANNING, Christopher D., RAGHAVAN, Prabhakar, SCHÜTZE, Hinrich. **Introduction to information retrieval.** Cambridge University Press, pp. 117-119, 2008.
[4] ALLAHYARI, Mehdi et al. **A brief survey of text mining:** Classification, clustering and extraction techniques. arXiv preprint arXiv:1707.02919, 2017.