

UTILIZAÇÃO DE MINERAÇÃO DE DADOS PARA CLASSIFICAÇÃO DE GÊNERO EM TEXTOS

Giovane Pace Daidone (IC) e Leandro Augusto da Silva (Orientador)

Apoio: PIBIC Mackenzie

RESUMO

A geração de dados tem chamado a atenção de pesquisadores não apenas pelo volume, mas também pela variedade. Nas redes sociais, tecnologia responsável pelo aumento significativo na produção de dados no mundo, tem-se uma variedade no formato da informação disponível, e em grande parte, em texto. No entanto, outras informações importantes para análises em um contexto mais amplo de aplicação, não estão disponíveis, por exemplo, o gênero do participante da rede social. Nessa ausência de informação que este trabalho de pesquisa se insere, através da representação dos dados por características capazes de descrever o texto pelo estilo de escrita do autor. O objetivo do trabalho é reconhecer o gênero do usuário da rede social. Os melhores resultados de classificação foram gerados pelo classificador *árvore de decisão*, analisados através dos resultados de acurácia para descritores textuais.

Palavras-chave: Arvore de decisão, análise de texto, classificação de dados.

ABSTRACT

The data generation has attracted the attention of researchers not only because the volume, but also by the data variety. In social networks, which is responsible for the significant increase in data production in the world, has a variety of information available in the format, mainly, in text. However, other important information for social network analysis are not available, for example, the gender of person. This lack of information is where this research is inserted, which the purposes are the data representation for characteristics that describe the text by the author's writing style. The objective is to recognize the user's gender social network. The best classification results were generated by the decision tree classifier, whose accuracy results for textual descriptors.

Keywords: decision tree, text analysis, gender in social networks.

1. INTRODUÇÃO

No início a escrita era feita através de desenhos, cada imagem significava algo e seu uso era bastante complicado. Esse sistema evoluiu para uma série de marcas conhecida como cuneiforme, escritas em argila e foram usadas amplamente no Oriente Médio. A escrita cuneiforme evoluiu para a ideográfica, que não utilizava apenas rabiscos associados a imagens, mas sim uma figura que representasse uma ideia, tornando-se depois uma convecção escrita, de onde veio o nosso alfabeto (criação no início do ano 2000 a.C.). Conforme a evolução da tecnologia, veio a escrita eletrônica, que não se limita aos textos, podendo ser palavras, imagens, sons, ações e outros (MEC, s.d.).

Hoje, na era da informação digital há o problema da sobrecarga de dados e a capacidade de analisá-los e entendê-los está muito além de conseguir armazenar, afinal a análise e entendimento auxiliam em processos decisórios e podem gerar um grande diferencial para diversos fins, procurando conhecimento (FAYYAD et. al., 1996; CARDOSO et. al., 2008). Com a chegada dos dispositivos móveis, essa situação se tornou ainda mais intensa.

A quantidade de dados disponível para análise e interpretação é muito maior que o uso efetivo dos mesmos em processos decisórios ou geração de conhecimento (FAYYAD et. al., 1996; CARDOSO et. al., 2008).

Uma das dificuldades para análise e interpretação pode ser a ausência de informações, principalmente de dados disponíveis em redes sociais. Adicionalmente, os dados disponíveis em redes sociais são passíveis de alguns problemas como veracidade de informações do tipo nome, gênero, idade etc. (CHENG et al., 2011). Sendo assim, problemas como estes de ausência ou veracidade de informações geram dificuldades na utilização dos dados em análises e abrem caminhos para diferentes abordagens de pesquisas. Exemplos de trabalhos possíveis nesse contexto relacionam-se à segurança de informação através do combate aos crimes cibernéticos, ao apoio em marketing para apresentar campanhas publicitárias aderente ao usuário, à temas relacionados com recuperação de informação para permitir explorações sobre os dados multimídia (texto, imagem e vídeo) e muitas outras aplicações (CHENG et al., 2011).

Tendo em vista os problemas acima apresentados e a abrangência de aplicações possíveis usando dados de redes sociais, este trabalho tem como objetivo principal a classificação de gênero de usuários da rede social *LinkedIn*, através da foto do perfil e texto descritivo de experiências profissionais. Como objetivo específico do trabalho, propõe-se verificar a eficiência da representação dos dados multimídia (imagens e textos), através de características visuais, textuais e combinação de ambas no reconhecimento do gênero. A

representação visual será feita usando técnicas capazes de descrever as imagens por cor, forma e textura. A representação textual, por outro lado, pretende explorar características que expressam o estilo de escrita dos autores por meio da extração de contagens de palavras, classes gramaticais, números, pontuações entre outras. Além das análises individuais dos descritores, pretende-se também verificar a possibilidade de combinação dos dois conjuntos de características, visuais e textuais. Os resultados deste trabalho, além do uso em classificação de gênero, ainda podem ser úteis como sistema de apoio a crimes cibernéticos, buscas por gêneros em redes sociais, sugestão de anúncios exclusivos para cada tipo gênero, apoio para identificação de autenticidade de documentos entre outros.

Tendo em vista as tecnologias de mineração de dados disponíveis e bases de dados de livre acesso, questiona-se o homem e a mulher usam estilos diferentes de linguagem e se isso é verídico, quais serão essas diferenças que podem identificar qual é o gênero do autor de um texto, usando a tarefa de classificação (CHENG et. al., 2011).

Além da introdução que tem como objetivo contextualizar o leitor, na seção 2, apresenta-se o embasamento teórico necessário para aprofundar o leitor no tema. A seção 3 informa ao leitor as metodologias usadas na pesquisa. Os resultados e análises experimentais do presente trabalho estão dispostos na seção 4 e a análise dos resultados na seção 5, a conclusão na seção 6 e por fim, na seção 7 estão as referências bibliográficas utilizadas para a construção do trabalho.

2. REFERENCIAL TEÓRICO

A classificação de dados é uma tarefa da mineração de dados em que, a partir de um modelo preditivo treinado, pode-se reconhecer automaticamente a classe do objeto em estudo (TAN et. al., 2005). É uma tarefa intuitiva ao ser humano, pois é natural a tentativa de classificar os objetos ou informações disponíveis com base no conhecimento que está intrínseco ou que foi constituído ao longo da vida (CARVALHO, 2005).

Um dos principais algoritmos de classificação de dados disponíveis na literatura é o *k-NN* (em inglês, *Nearest Neighbors*), que recebe um objeto descrito por um conjunto de atributos (ou características) sem a variável classificatória e, o algoritmo, através de medida de similaridade, compara esse objeto sem classe com os atributos de um conjunto de dados classificado. E assim, encontra-se o *k* (valor inteiro maior ou igual a 1) vizinho mais semelhante, atribuindo ao objeto em análise, a classe do *k* mais próximo (para $k > 1$ a decisão é pela classe majoritária) (COVER; HART, 1967).

Árvore de decisão é outra da técnica de classificação de dados que é gerada por um processo de estruturação dos atributos do conjunto de dados para algum critério, como por exemplo pelo ganho de informação. Nessa árvore os nós são atributos, os ramos os valores possíveis e as folhas são as classes. Em um processo de classificação, um caminho da árvore é traçado, a partir do nó raiz, até o nó folha (ou terminal) que resulta na previsão de classe (SAFAVIAN; LANDGREBEL, 1991). Para a classificação dos dados multimídias utilizados neste trabalho, é necessário transformá-los em instâncias numéricas estruturadas com base em suas características para que as dimensões, formatos e outros detalhes que as tornam diferentes entre si auxiliem no reconhecimento de objetos. (HARZALLAH et. al., 2009).

Para o caso de dados no formato texto, existe a diferença na quantidade de caracteres e diferença de fonte (família, tamanho e cor) e, portanto, deve-se extrair as suas principais características que os deixem em dimensões equivalentes para análise, sem a perda de informação discriminatória.

Nas seções seguintes serão apresentados mais detalhes sobre as técnicas de extração de características para texto abordadas neste trabalho.

A estruturação de documentos textuais é feita por um modelo de representação que transforma as palavras de um documento em um vocabulário. O resultado disso é chamado de *corpus* e, a partir desse, é que as características do texto são extraídas através de atribuição de um peso a cada item do vocabulário. Esse processo ocorre, basicamente, da seguinte maneira (GÖKER; DAVIES, 2009):

- a) *Tokenization*: a partir do caractere espaço, as palavras de um documento são separadas em *tokens*. Os caracteres especiais como vírgulas (","), pontos de exclamação ("!") e de interrogação ("?"), entre outros são removidos, assim como números;
- b) *Stopwords*: palavras como preposições, pronomes, artigos, advérbios que são comuns a ambos os vocabulários são removidas do processo;
- c) *Stemming*: as palavras resultantes das etapas anteriores passam por uma normalização ortográfica para que sejam reduzidas ao radical. Este processo é importante, pois permite que palavras com o mesmo radical sejam consideradas como semelhantes.

Ao final deste processo de transformar um texto em um *corpus*, as palavras são armazenadas em seus respectivos vocabulários. A representação dos documentos é feita usando um modelo chamado espaço vetorial (GÖKER; DAVIES, 2009), que busca representá-lo em forma de vetor, composto por termos (palavras) representados por um peso ponderado. Uma das formas de se fazer a ponderação é combinando a frequência que um

termo aparece (*tf* do inglês, *term frequency*) com a frequência invertida do documento (*idf* do inglês, *inverse document frequency*). A combinação de *tf* x *idf* define a importância (peso) de um termo dentro do documento (GÖKER; DAVIES, 2009).

Assim, depois de calculado o peso de cada termo, tem-se como resultado uma base de dados numérica formada por vetores de documentos chamada de *Bag of Words* (BoW). No entanto, o BoW é uma matriz esparsa e, dependendo da sua utilização, a representação é ampla, não permitindo assim capturar as características principais do texto.

Alternativamente, apresenta-se neste trabalho a extração de metadados dos textos. Os metadados são representações do texto através de contagens de cada tipo de palavra que pertence ou não a uma classe gramatical. Por exemplo, captura da quantidade de verbos, substantivos, adjetivos, números e entre outros. Esta captura é feita a partir de expressões regulares, sendo elas objetos que descrevem padrões de caracteres. As expressões regulares são usadas para executar reconhecimento de padrões e funções "procurar e substituir" no texto (W3C, 2015).

As expressões regulares criadas visam encontrar características do português brasileiro. Segue a lista com as expressões utilizadas na Tabela 1. Nesta tabela apresenta-se também um exemplo de aplicação, com a palavra a ser contada destacada.

Tabela 1 - Tabela de expressões regulares utilizadas para obter metadados.

Expressão Regular	Padrão identificado	Exemplo
(o)(os)(a)(as)+	Artigos Definidos	Utilizamos <u>a</u> ferramenta
(um)(uns)(uma)(umas)+	Artigos Indefinidos	Comemos <u>umas</u> bolachas
(€)(\\$(¥)(R\\$))(reais)(reais)(reais\.)(dolar)(dolar)(dolar\.)(euro)(euro)(euro\.)(iene)(iene)(iene\.))+	Indicadores monetários	Esse saco de batatas custou 25 <u>reais</u>
[\t]	Espaço em branco e tabulação:	Meu_filho_adora_jogar
[a-z]	Letras em caixa baixa	APENAS MIN <u>us</u> culas
[A-Z]	Letras em caixa alta	<u>APENAS</u> MINusculas
[0-9]	Números	Temos <u>3</u> aulas hoje
[+-.!@#%&*() \\ <>"]	Caracteres especiais	Adoro emojis <u>^^</u>
[.,!,:?]	Pontuação	Onde comprou essa blusa <u>?</u>
[âãäåöðõöëéèú]	Letras acentuadas	<u>Não</u> trabalharei amanhã
\w+([aei]lpo)r\b	Verbos no infinito	Minha namorada adora <u>conversar</u>

Fonte: Elaborada pelos autores

3. METODOLOGIA

O problema da identificação de texto pode ser tratado como binário, já que existem duas classes (feminino e masculino) e para chegar ao objetivo deste projeto, serão selecionadas bases de dados com textos previamente anotados. Serão estudadas as características existentes entre a escrita dos gêneros, com experimentos para verificar a melhor combinação.

Para classificação de dados multimídias como os que serão abordados neste trabalho, é necessário um processo para transformar esses dados em uma instância numérica estruturada. Isto é, consiste em extrair características do texto que sejam capazes de descrever o conteúdo do mesmo, ou seja, que o represente.

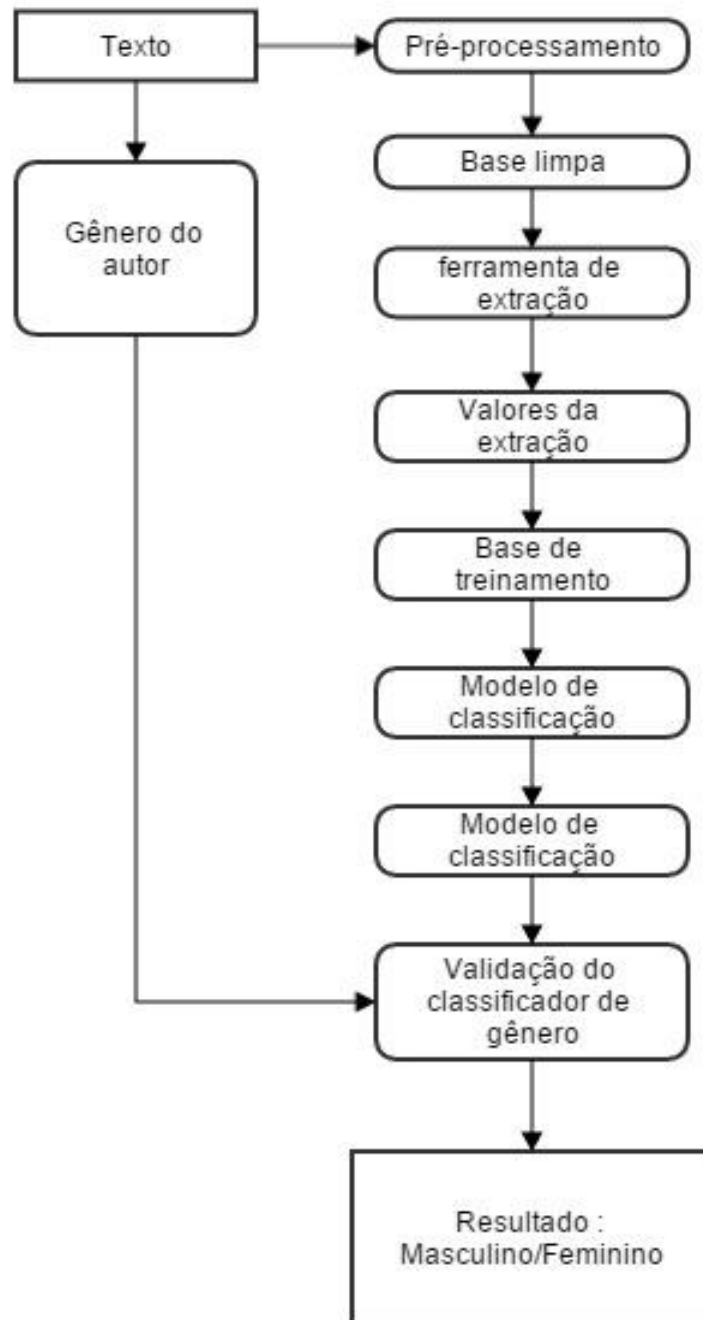
Estruturação de dados textuais

Para treinar a base, serão utilizadas expressões regulares empregadas a fim de encontrar os metadados dos textos de resumo dos perfis do *Linkedin*. Será realizada a contagem dos metadados encontrados em cada texto, a fim de encontrar disparidades e padrões entre os gêneros de masculino e feminino.

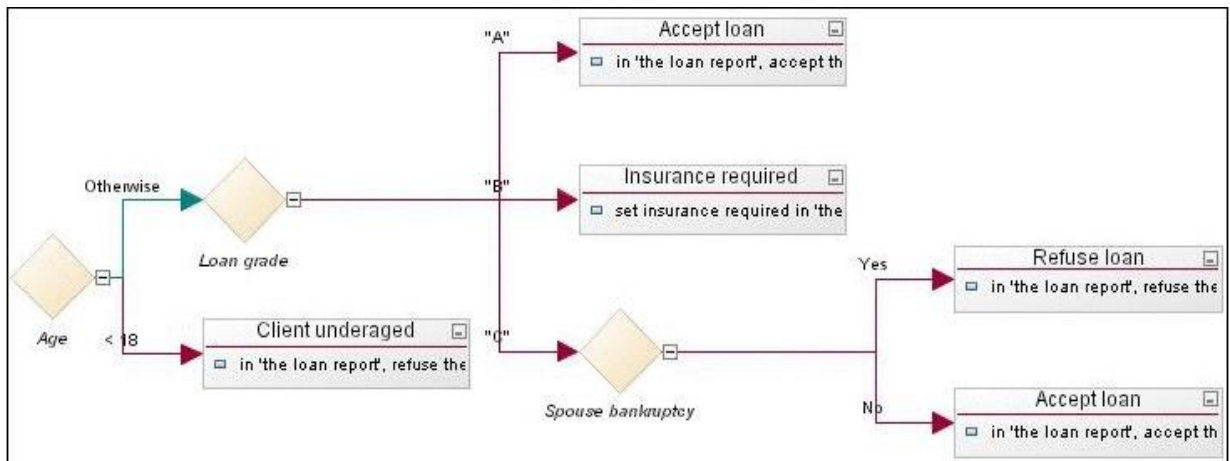
Abaixo, a função em *Java* que encontra a quantidade de ocorrências de uma expressão regular em um conjunto de caracteres, que serão utilizadas para a análise dos metadados de um texto.

Alguns exemplos de expressões utilizadas são: artigo definido, artigo indefinido, verbo, valores monetários, entre outros. Os dados encontrados serão exportados em um arquivo *Comma-Separated Values (CSV)*, para serem submetidos aos classificadores.

Figura 1 : Processo de identificação de texto



Fonte: Elaborada pelos autores

Figura 2 : Exemplo do algoritmo de árvore de decisão

Fonte: IBM(2015)

Quadro 1 - Código java para contagem de palavras

```

public int achaQtd(String texto,String regex){
    //Cria um padrão de caracteres
    Pattern pattern = Pattern.compile(regex);
    //Cria a classe matcher, que encontra o padrão em um texto
    Matcher matcher = pattern.matcher(texto);
    int count = 0;
    //Loop realizando a contagem
    while (matcher.find())
        count++;
    return count;
}

```

Fonte: Elaborado pelos autores

Classificação dos Dados

Os arquivos CSV contendo os resultados dos descritores visuais e das expressões regulares serão submetidos a dois métodos de classificação, sendo eles o *k*-NN e a árvore de decisão. Foram escolhidos dois métodos para fins comparativos de ambos os resultados. Os arquivos foram processados na linguagem *R*, que possui bibliotecas nativas com os métodos de classificação utilizados. O código gerado retorna uma matriz de confusão e a acurácia, resultantes da aplicação da técnica *leave-one-out*.

A técnica *leave-one-out* é uma das muitas que existem para realizar a avaliação dos modelos de classificação. Ela consiste em retirar um exemplar da base de treinamento e aplicar o modelo de classificação nesse exemplar e finalmente comparar o resultado obtido com o esperado. O processo é repetido para todos os exemplares da base (TAN et. al., 2005).

Os resultados do método *leave-one-out* são compilados em uma matriz denominada matriz de confusão. A matriz de confusão é o encontro dos valores verdadeiros para os exemplares, que são expressos pelas linhas, e dos resultados que foram obtidos através do modelo de classificação, sendo estes as colunas, conforme a Tabela 3. O encontro das colunas e das linhas resultam nos seguintes coeficientes (SILVA, 2015):

- a) Verdadeiro Positivo (VP): Quando o classificador encontra a classe correta para a classe positivo;
- a) Verdadeiro Negativo (VN): Quando o classificador encontra a classe errada, ou seja, classifica como negativo a classe positivo;
- a) Falso Positivo (FP): Quando o classificador encontra a classe correta para a classe negativo;
- a) Falso Negativo (FN): Quando o classificador encontra a classe errada, ou seja, classifica como positivo a classe negativo;
- a)

Neste trabalho, a classe denominada como positivo, representa o gênero masculino e a classe negativo, o gênero feminino.

Tabela 2 - Matriz de Confusão.

		Classe Prevista	
		positivo	negativo
Classe Verdadeira	positivo	VP	FN
	negativo	FP	VN

Fonte: SILVA (2015)

A partir dos resultados da matriz de confusão é possível extrair outras medidas com base em seus resultados, como a sensibilidade e especificidade (SILVA, 2015).

$$\text{Sensibilidade} = \frac{VP}{VP + FN}$$

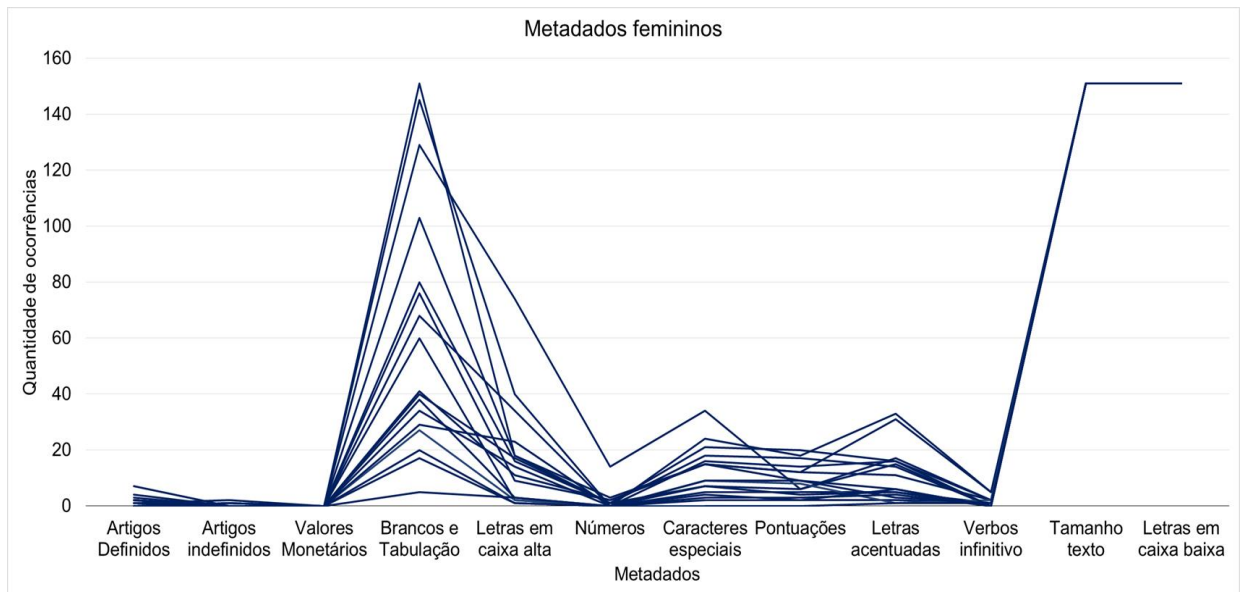
$$\text{Especificidade} = \frac{VN}{VN + FP}$$

Uma das características da matriz de confusão é permitir que seus resultados sejam expressos pela acurácia. A acurácia de um classificador é a probabilidade de acertar a classificação de uma linha escolhida aleatoriamente em um conjunto de dados pela qual foi juntamente treinada (KOHAVI, 1995). A acurácia é medida por meio da matriz de confusão a partir da seguinte fórmula (SILVA, 2015):

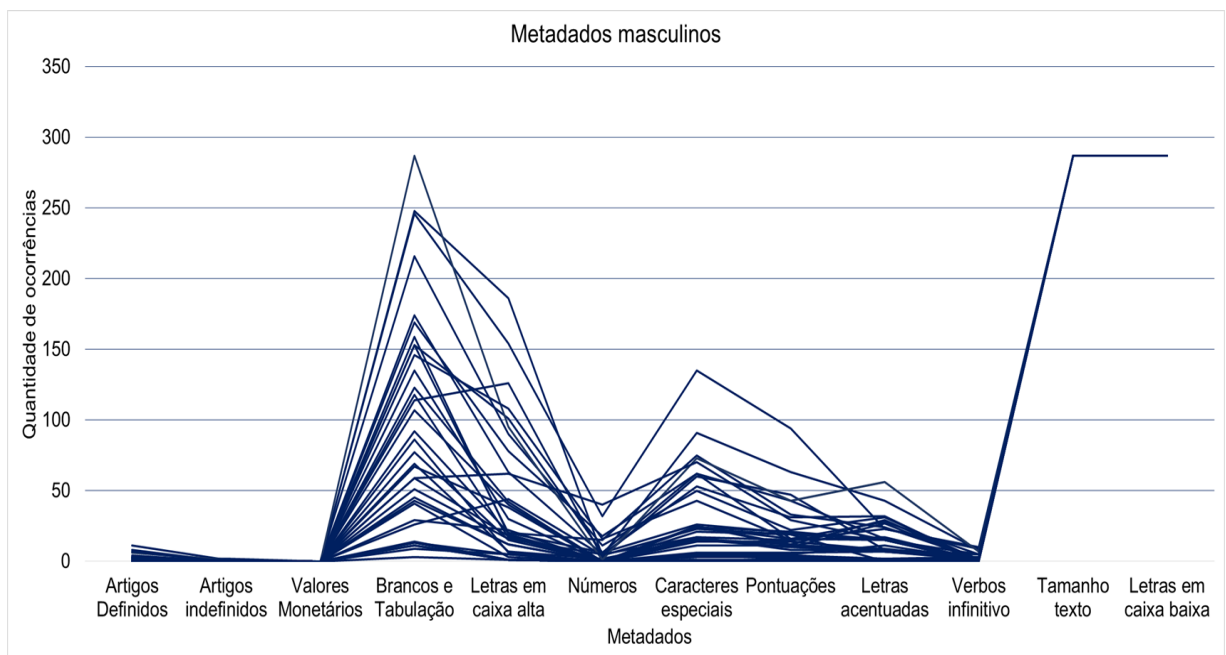
$$\text{Acurácia} = \frac{VP + VN}{VP + FN + FP + VN}$$

4. RESULTADOS E ANÁLISES EXPERIMENTAIS

Serão apresentados o Gráfico 9 e Gráfico 10, contendo os valores das expressões regulares para análise dos pontos de maior diferença, e melhor percepção de resultados na base de treinamento. Os valores de tamanho de texto e letras em caixa baixa, foram alterados para manter uma boa proporção do gráfico, por serem bem maiores que os outros metadados.

Gráfico 1 - Valores dos metadados utilizados em linha para o gênero feminino.

Fonte: Elaborada pelos autores

Gráfico 2 - Valores dos metadados utilizados em linha para o gênero masculino.

Fonte: Elaborada pelos autores

No primeiro teste realizado, foi utilizado o método de classificação *k-NN* para os metadados encontrados no texto e no método de referência, o *Bag of Words*, como mostra a Tabela 5.

Tabela 3 - Tabela de expressões regulares utilizadas para obter metadados

		Acurácia	Sensitividade	Especificidade
Metadados textuais	<i>k-NN</i>	53,85%	68,29%	33,33%
BoW	<i>k-NN</i>	61,54%	68,89%	36,36%

Fonte: Elaborada pelos autores

Outro teste foi realizado, dessa vez separando as expressões regulares, por opções mais granularizadas, ou seja, separamos as ocasiões dos conjuntos de artigos indefinidos, artigos definidos e pontuações, além de usar também o algoritmo de árvore de decisão, assim como mostra a Tabela 6.

Tabela 4 - Tabela de expressões regulares utilizadas para obter metadados, pela segunda vez, utilizando mais granularidade.

		Acurácia	Sensitividade	Especificidade
Textuais da segunda versão	<i>k-NN</i>	51,92%	62,86%	29,41%
	Árvore de decisão	61,54%	62,86%	58,82%

Fonte: Elaborada pelos autores

5. CONCLUSÃO E TRABALHOS FUTUROS

A combinação dos descritores textuais para esta base de dados, geraram uma acurácia de 61,54%, o que levando em consideração que o resultado padrão é binário (feminino ou masculino) a um avanço com uma maior probabilidade de acerto.

Como possibilidade para trabalhos futuros, para descritores textuais, os metadados obtidos do texto poderiam ser expandidos e melhorados, criando outras expressões regulares que possam ter relevância, a fim de conseguir maior acurácia, já que são simples e de baixo custo computacional. O estudo com metadados pode abrir desdobramento para novas pesquisas, como utilizar o texto para descobrir efetivamente uma pessoa, ou seja, as características que uma pessoa usa na hora de escrever, analisando o estilo de escrita para ajudar na descoberta de autores de códigos fonte e direitos autorais.

Para a análise de texto, podem ser usados outros descritores ou até mesmo outras combinações dos mesmos utilizados neste trabalho para verificar se é possível melhorar o

resultado, uma vez que a classificação de dados estruturados e não estruturados feita de maneira conjunta pode abrir novas abordagens de pesquisa.

REFERÊNCIAS

A. Herdağdelen, M. Baroni. **Stereotypical gender actions can be extracted from Web text**. Journal of the American Society for Information Science and Technology, Volume 62, edição 9, pages 1741–1749. Cambridge, Estados Unidos, 2011.

A. Herdağdelen. **Gender differences in tweets: A corpus-based analysis of Twitter**. Turunç Workshop on Complex Systems, Marmaris, Turquia, 2010.

BALLARD, Dana H.; BROWN, Christopher M.. **Computer Vision**. Nova Jersey: Prentice-hall, 1982.

BAMMAN, David; EISENSTEIN, Jacob; SCHNOEBELEN, Tyler. **Gender identity and lexical variation in social media**. 2. ed. Nova Iorque: Journal Of Sociolinguistics, 2014. 25 p. Disponível em: <<http://onlinelibrary.wiley.com/doi/10.1111/josl.12080/full>>. Acesso em: 08 jun. 2015.

BESL, PAUL J.; JAIN, RAMESH C.. In: **Computer Vision and Pattern Recognition Conference**, 1985, São Francisco, California. Proceedings. IEEE 1985. P 430-451.

CANNY, John. A Computational Approach to Edge Detection. **Ieee Transactions On Pattern Analysis And Machine Intelligence**. Cambridge, p. 679-698. nov. 1986.

CARDOSO, O.; MACHADO, R. Gestão do conhecimento usando data mining: estudo de caso na Universidade Federal de Lavras. **Revista de administração Pública**, Rio de Janeiro, Vol 42, 495-528, Junho. 2008.

CARVALHO, Luís Alfredo Vidal de. **Datamining – A Mineração de Dados no**

Marketing, Medicina, Economia, Engenharia e Administração. Rio de Janeiro: Editora Ciência Moderna LTDA, 2005.

Cheng, N. Author gender identification from text. **Digital Investigation**, Hoboken, Estados Unidos, 2011, Vol. 8, 78-88, 2011.

CHRISTMANN, Raul Udo. **Estatística aplicada**. São Paulo: Edgard Blucher, 1978. 135p.

COVER, T. M.; HART, P. E.. Nearest Neighbor Pattern Classification. **Ieee Transactions On Information Theory**, New Jersey, v. 13, n. 1, p.21-27, dez. 1967.

DINIZ, C.A. & LOUZADA NETO, F. **Data Mining: uma introdução**. São Carlos: Associação Brasileira de Estatística, 2000.

D. HAND, H. Mannila, P. Smith. **Principles of Data Mining**. MIT Press, 1998.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth P.. The KDD Process for Extracting Useful Knowledge from Volumes of Data, **COMMUNICATIONS OF THE ACM**. [S. l.] Novembro de 1996/Vol. 39, No. 11.

IBM. **Árvores de decisão**. Acesso em 22 fev. 2015. Disponível em: <http://www.ibm.com/support/knowledgecenter/SSQP76_8.0.0/com.ibm.wodm.dcenter.bu.rules/shared_dtree_topics/con_decision_tree.html?lang=pt-br>

GÖKER, A., & DAVIES, J. (2009). **Information retrieval: searching in the 21st century**. Chichester U.K.: John Wiley & Sons, Inc.p.6-7.

JARGAS, Aurelio Marinho. **Expressões Regulares - Uma abordagem divertida**. São Paulo: Editora Novatec, 2012.

KOHAVER, Ron, 1995. **Study of a cross-validation and bootstrap for accuracy estimation and model selection**. Disponível em: <<http://web.cs.iastate.edu/~jtian/cs573/Papers/Kohavi-IJCAI-95.pdf>>. Acesso realizado em 18/10/2013.

MÄKINEN, Erno; RAISAMO, Roope. An experimental comparison of gender classification methods. **Pattern Recognition Letters**. Tampere, p. 1544-1556. abr. 2008.

Ministério da educação. A história da escrita. Acessado em 15/09/2014 as 8:40. Disponível em:<http://webeduc.mec.gov.br/midiaseducacao/material/impresso/imp_basico/e1_assuntos_a1.html>

O'Hara, S. & Draper, B.A., 2011. **Introduction to the Bag of Features Paradigm for Image Classification and Retrieval**. Disponível em: <<http://arxiv.org/abs/1101.3354>>. Acesso realizado em 08/06/2013.

SAFAVIAN SR, Landgrebe D. **A survey of decision tree classifier methodology**. Maio de 1991.

SCHIELE, Bernt; CROWLEY, James L.. Object recognition using multidimensional receptive field histograms. In: ECCV '96 PROCEEDINGS OF THE 4TH EUROPEAN CONFERENCE ON COMPUTER VISION, 4., 1996, Grenobla. **Proceedings...** . Grenobla: Springer Berlin Heidelberg, 1996. v. 1, p. 610 - 619.

SILVA, Leandro Augusto da. **Mineração de Dados: Uma abordagem introdutória e ilustrada**. São Paulo: Mackenzie, 2015. 153 p.

TAN, Pang-ning; STEINBACH, Michael; KUMAR, Vipin. **Introduction to Data Mining**. Boston: Pearson, 2005.

WANG, Li; HE, Dong-chen. TEXTURE CLASSIFICATION USING TEXTURE SPECTRUM. **Pattern Recognition**. Grã-bretanha, p. 905-910. 1 jun. 1989.

W3C. **JavaScript RegExp Reference**. 2015. Disponível em: <http://www.w3schools.com/jsref/jsref_obj_regexp.asp>. Acesso em: 16 jan. 2015.

Contatos: giovanepace@gmail.com e prof.leandro.augusto@mackenzie.br