Investigating the Influence of Groups of Variables on the Task of Predicting the Age of an Author in Blog Posts

R. Neto, R. Ribeiro, and A.Emília

***Abstract*—The identification of the profile of users from texts on the Internet is a relevant task in the context of today's society. This activity is known in the literature as Author Profiling. Among the essential characteristics to be deduced in this task is the age. This feature is paramount, for example, for the identification of potential sexual predators in environments targeted for children. However, one of the issues faced in resolving this problem is the identification of which variables should be taken into account to address this problem. Thus, this article aims to identify which variables are relevant in building a data mining solution to infer a user's age from a text on the Internet. An experimental study was carried out in a database of a prestigious international competition, considered a benchmarking of the area, to validate this work. The results showed that there is a difference between the possibilities of variables that can be constructed to solve this problem and justifies the importance of each variable group for this purpose. The main contribution of this study was to find different relevance among groups of variables previously mentioned in the literature.**

***Index Terms*—Author Profiling, Age Identification, Text Mining, Data Mining.**

I. Introdução

A Caracterização de autoria fornece uma grande aplicabilidade, especialmente no contexto da internet. Como exemplo, podemos citar as aplicações forenses. Perfis de autores de mensagens com conteúdo criminoso podem ser montados, auxiliando na sua identificação [1]. A idade de pessoas em ambientes voltados para o público infantil pode ser estimada a fim de detectar potenciais predadores sexuais. Além disso, o *Marketing* pode se beneficiar dos avanços na área. Empresas podem ligar as opiniões encontradas em redes sociais e sites de opinião à determinados perfis, permitindo uma compreensão melhor de seus consumidores [2]. Propagandas podem ser melhor direcionadas à um público específico a partir de um maior conhecimento das características de seu público-alvo [3]. A caracterização de autoria é um campo de pesquisa da Mineração de Texto, subárea de Mineração de Dados que tem se tornado popular desde o fim do século XX. Ela envolve a extração de conhecimento previamente desconhecido de textos e difere da Mineração de Dados tradicional porque seus dados não estão estruturados em bancos de dados ou outras fontes facilmente interpretáveis por computadores [4]. Eles precisam ser processados de modo a gerar variáveis que representem informação. No entanto, surge uma questão: "quais variáveis construir?". O principal desafio na proposição de uma solução de caracterização de autoria é justamente a construção dessas variáveis [5]. Este artigo identificou os principais grupos de variáveis que podem ser construídos e analisou sua eficiência. Um estudo experimental foi realizado para identificar a relevância destes grupos de variáveis. O estudo utilizou a base de dados de uma importante competição internacional, a PAN@CLEF [6]. Os resultados mostram que, quando considerados isoladamente, os grupos de variáveis apresentam desempenho diferente. A principal contribuição deste trabalho foi encontrar diferente relevância entre grupos de variáveis previamente apontados na literatura. Além disso, os resultados evidenciam que um maior poder preditivo é alcançado com a combinação das variáveis. O restante do artigo está dividido como segue. A seção 2 apresenta a definição do problema. A seção 3 detalha os principais grupos de variáveis utilizados na literatura. A seção 4 apresenta os trabalhos relacionados. A seção 5 exibe a metodologia experimental adotada no estudo. A seção 6 apresenta os resultados obtidos para validação do estudo. Por fim, a seção 7 conclui o trabalho e propõe trabalhos futuros.

Rosalvo Neto, Rodrigo Ribeiro Oliveira and Ana Emília were with the Department of Computer Engineering, Universidade Federal do Vale do São Francisco (UNIVASF), Juazeiro, Bahia, Brasil e-mail: (rosalvo.oliveira, rodrigo.oliveira, ana.queiroz) @univasf.edu.br.

II. Abordagem do Problema

A classificação de textos em relação à autoria é um problema comum em Mineração de Texto [7]. Ela envolve um conjunto previamente definido de autores e um conjunto de *corpus* associado a cada autor. Uma definição de *corpus* é "uma coletânea de textos em linguagem natural, escritos ou falados, geralmente armazenados de forma organizada e informada, além de serem digitalizados a fim de que possam ser lidos por computador" [8]. A tarefa consiste em criar um classificador capaz de predizer a autoria de textos anônimos a partir autores previamente informados. Neste trabalho, foi considerada uma variação deste problema, a caracterização de autoria. Para este problema, não são disponibilizados previamente os autores dos *corpus*. O objetivo é inferir informações relevantes sobre o autor como, por exemplo: sua idade [9]. Por consequência, a tarefa consiste em classificar um ou mais textos em categorias que indicam perfis de autoria, como por exemplo, o grupo de idade do autor.

Estes grupos são construídos a partir da observação de que classes diferentes de pessoas se expressam de modo peculiar. Para capturar tais diferenças, são construídos, como no problema original, variáveis a partir do texto, que são utilizados

para alimentar algoritmos de aprendizagem de máquina que classificarão os textos de autoria desconhecida. De acordo com [9], o processo de caracterização de autoria envolve quatro etapas:

- Os dados de treinamento são definidos a partir um *corpus*, previamente rotulado de acordo com as características utilizadas;
- Cada texto no *corpus* é processado a fim de gerar um vetor contendo um conjunto de variáveis para diferenciar as classes a que os autores pertencem;
- Métodos de aprendizagem de máquina são aplicados ao conjunto de variáveis previamente identificado para gerar um classificador;
- O classificador gerado é usado para definir as características dos dados de autoria desconhecida.

Um dos maiores desafios neste processo é saber quais variáveis devem ser construídas. O processo de construção de variáveis é um dos mais antigos e ainda desafiadores problemas quando são propostas soluções de mineração de dados [10]. De acordo com [5], a melhor maneira de construir variáveis é manualmente, baseado no entendimento do problema de aprendizagem. De uma forma geral, a tarefa de construção de variáveis é muito mais dependente do domínio do que a construção de um classificador, por isso o conhecimento do domínio é relevante. Neste contexto, na hipótese dos profissionais ou pesquisadores não terem conhecimento do domínio de classificação de autoria, é de extrema relevância ter um direcionamento de quais variáveis podem ser consideradas neste processo.

III. Grupos de Variáveis

Diversas variáveis podem ser construídas para caracterização de autoria. A seguir serão descritos os nove grupos de variáveis mais utilizadas de acordo com o levantamento bibliográfico realizado neste trabalho.

- **Medidas de complexidade**: um potencial discriminador de características de autoria está na complexidade textual de cada classe [11]. Desta forma, variáveis que expressam a complexidade do texto podem ser calculadas, entre elas podemos citar: o número absoluto de caracteres, palavras e sentenças [6]. Outra variável que pode ser calculada é a diversidade, que correspondente à razão entre o número de palavras distintas e o total de palavras no texto. Além dessas, outras variáveis mais elaboradas também podem ser utilizadas como, por exemplo: o Teste de Leitura Flesch (Flesch Reading Ease, FRE), que a partir de uma expressão, indica o quão fácil um texto pode ser lido; e o Teste de Classificação Flesch-Kincaid (Flesch-Kincaid Grade Level, FKGL), que liga a complexidade de um texto à uma série do Sistema Educacional Americano [3].
- **Palavras-função**: apesar de o número grande de palavras que constitui o léxico de uma língua, o discurso falado e escrito é composto de um conjunto comparativamente pequeno. Por exemplo, empresas especializados na área apontam que na língua inglesa existem 2.000 palavras mais freqüentes, que representam 80-85% das palavras em textos escritos não especializados e cerca de 90-95% em fala coloquial (linguagem falada informal) [1]. Destas palavras mais frequentes, existem um grupo chamado de palavras-função, do inglês *function words*. Este grupo de palavras seria o equivalente em língua Portuguesa aos termos acessórios, que são palavras que podem ser retiradas de uma sentença mantendo seu significado, diferente dos verbos (termos essenciais). A utilização das palavras-função é feita a partir da contagem de ocorrências de palavras-função de uma lista pré-estabelecida com frequência elevada nos textos como, por exemplo, pronomes, preposições, verbos modais, conjunções, entre outros [12]. O motivo para usar palavras função é que não é esperado que suas freqüências variem muito com o tópico do texto e, portanto, é possível reconhecer textos do mesmo autor ou classe de autores em diferentes tópicos [13].
- **Classes gramaticais**: o desenvolvimento da área de Processamento de Linguagem Natural tornou possível identificar e usar as classes gramaticais das palavras dos textos, como verbos, pronomes, adjetivos, artigos entre outros. Diferente da abordagem de palavras-função, não é realizada a contagem de cada palavra presente em uma lista de palavras-função, mas sim o total de palavras em cada classe gramatical [11]. Desta forma, é calculada a frequência relativa de cada classe em relação ao total de palavras no texto [3]. Essa abordagem é conhecida como parte do discurso [6], do inglês *Part of Speech* (POS).
- **Taxonomia Léxica**: uma abordagem de construção de variáveis é combinar as classes gramáticas com palavras-função em uma taxonomia. Essas taxonomias são representadas como árvores, onde o nó raiz representa as classes gramaticais e os nós filhos representam o significado deste grupo de palavras [3]. Durante o levantamento bibliográfico realizado neste trabalho, não foram encontradas taxonomias funcionais de domínio público disponíveis em *frameworks*.
- **Emoções**: associar emoções a textos é uma outra abordagem para construção de variáveis [6]. Ela consiste em contar quantas palavras estão associadas a uma lista de emoções disponíveis em um dicionário que associa diversas palavras com suas emoções [14]. É importante destacar que esta abordagem de construção de variáveis é diferente de Análise de Sentimentos, que é um caso especial de mineração de texto focado em classificar um texto em dois sentimentos: positivo e negativo.
- **Corretude**: capturar informação da corretude ortográfica de um texto é outra abordagem de construção de variáveis [15]. Como exemplo podemos citar: a razão entre as palavras encontradas em um dicionário e o total de palavras do texto, número de vogais repetidas e número de sinais de pontuação repetidos, entre outros.
- **Frequência de Palavras**: Esta abordagem cria uma lista de variáveis que contém as palavras mais frequentes em todos os *corpus* disponíveis [3]. O conteúdo de cada variável é a frequência de cada palavra dessa lista no *corpus* corrente [13]. Esta abordagem é conhecida

[1] https://www.sequencepublishing.com/1/academic/academic.html

em inglês como *bag of words* [6]. Uma variação desta abordagem é não utilizar a lista com as palavras mais frequentes, e sim utilizar uma lista com palavras que expressam informações específicas de um determinado domínio como, por exemplo: palavras relacionadas a trabalho, vida pessoal, relacionamentos entre outros [14].

- **N-Gram**: considera o número de vezes em que n palavras aparecem juntos em um documento [12]. Desta forma, combinações de dois termos são 2-grams, de três termos 3-grams, e assim por diante [3]. A abordagem de *bag of words* pode ser considerada a utilização de 1-gram [6].
- **Recuperação da Informação (do inglês *Information Retrieval*)**: essa abordagem é inspirada na área de Recuperação da Informação. Esta área de pesquisa tem como objetivo desenvolver algoritmos para procurar referências de texto associadas com um assunto. As preocupações iniciais dessa área eram: como indexar documentos e como recuperá-los. Dentro da tarefa de caracterização de autoria, essa abordagem constrói variáveis a partir de algoritmos de Recuperação da Informação e os transformam em atributos de entrada [15].

IV. Trabalhos Relacionados

Diversos trabalhos são encontrados na literatura sobre caracterização de autoria de texto. O objetivo desta seção é ilustrar a diversidade de variáveis que é utilizada nesta tarefa, e evidenciar que a escolha das variáveis é diversificada em cada estudo encontrado.

Em [12], os autores abordam a caracterização de autoria utilizando um *corpus* de documentos originários do *British National Corpus* (BNC), os quais apresentavam previamente definidos tanto autor e gênero quanto as classes gramaticais de todas as palavras. Os grupos de variáveis utilizados foram palavras-função, classes gramaticais e N-grams. Os resultados do estudo mostraram que determinadores e quantificadores caracterizaram melhor, autoria masculina, enquanto que pronomes pessoais indicaram autoria feminina.

Em [3], os autores investigam a detecção de gênero e idade utilizando um conjunto de blogs na língua inglesa. Os grupos de variáveis utilizadas foram taxonomias léxicas, frequência de palavras, N-grams e palavras funções. Como resultado, foi visto que determinadores, preposições e palavras relacionadas à tecnologia identificaram melhor, homens; enquanto que pronomes e palavras relacionadas a vida pessoal e a relacionamentos, mulheres. Quanto à idade, contrações sem apóstrofos e palavras relacionadas a escola e a humor indicaram idade mais jovem, já determinadores, preposições e palavras relacionadas a trabalho, a vida social e família, idade mais velha.

Em [13], os autores analisam o problema da detecção de gênero, idade, localização (Norte ou Sul do Vietnã) e profissão (estudante, cantor, modelo) utilizando blogs no idioma vietnamita. Os grupos de variáveis utilizados foram palavras-função, classes gramaticais e frequência de palavras como, por exemplo, tópico que associa uma palavra a um grupo de conteúdo como educação, saúde, economia entre outros. O estudo destaca que as variáveis relacionadas a palavras foram melhores do que as variáveis relacionadas com caracteres.

TABELA I
Frequência dos grupos de variáveis nos trabalhos relacionados. Legenda: A) Medidas de Complexidade, B) Palavras-Função, C) Classes Gramaticais, D) Taxonomia Léxica, E) Emoções, F) Corretude, G) Frequência de Palavras, H) N-Gram e I) Recuperação da Informação

	Grupo de Variáveis								
Artigo	A	B	C	D	E	F	G	H	I
[12]		X	X					X	
[3]		X		X			X	X	
[13]		X	X				X		
[11]	X		X						
[6]	X		X		X		X	X	
[15]	X				X	X			X
[14]					X		X	X	

Em [11], os autores abordam a classificação de gênero em bases de dados de resenhas de cinema do site *Internet Movie Database*. Os grupos de variáveis utilizados foram classes gramaticais, complexidade e variáveis propostas pelos autores como, por exemplo, dados acerca da resenha como atenção que ele recebeu da comunidade, avaliação dada ao filme pelo autor e riqueza de vocabulário. O resultado do estudo mostrou que as variáveis relacionadas a riqueza de vocabulário marcaram a autoria feminina, enquanto que o uso da terceira pessoa marcaram a autoria masculina.

Em [6], os autores investigam a detecção de gênero e idade utilizando um conjunto de blogs na língua inglesa. Os grupos de variáveis utilizados foram medidas de complexidade, classes gramaticais, N-Gram, emoções e algumas variáveis propostas pelos autores como frequências de palavras específicas, no entanto, os autores não fazem avaliações sobre as contribuições de cada grupo de variáveis.

Em [15], os autores analisam a detecção de gênero e idade utilizando um conjunto de blogs na língua inglesa. Os grupos de variáveis utilizados foram medidas de complexidade, emoções, corretude e recuperação da informação. O resultado do estudo mostrou que as variáveis de Recuperação da Informação foram as que proporcionaram maior poder discriminatório tanto para a tarefa de detecção de gênero como idade. Os autores destacam ainda que os resultados para o grupo de variáveis de análise de sentimento e corretude foram os piores.

Em [14], os autores identificam gênero em textos arábicos, usando um conjunto de dados extraído manualmente de sites como: alrai.com, addustour.com e sawaleif.com. O conteúdo desses sites abrange vários aspectos da vida. Não existindo, portanto, nenhum estilo ou formato preconcebido. Os grupos de variáveis utilizados foram N-Gram, frequência de palavras e emoções. O resultado do estudo mostrou que autores masculinos em textos arábicos tendem a escrever com maior número de sentenças, palavras e caracteres.

Para explicitar os elementos observados neste estudo, a Tabela I sintetiza os grupos de variáveis considerados em cada trabalho relacionado. Como pode ser visto, a utilização dos grupos de variáveis está bem dispersa. Usando como critério de ordenação a predominância de uso das variáveis,

TABELA II
FREQUÊNCIA DE POSTS NA LÍNGUA INGLESA DA PAN@CLEF 2013 POR CATEGORIAS DE IDADE

Categoria de Idade	Qtd. Treino	Qtd. Teste
10 anos	17200	1776
20 anos	85703	9170
30 anos	133508	14408

em primeiro lugar vêm as classes gramaticais, frequência de palavras e N-Gram, em segundo vêm as medidas de complexidade, palavras-função e emoções. Os grupos de variáveis que tiveram a menor representatividade foram corretude, taxonomia léxica e recuperação da informação. A principal justificativa para esta menor representatividade está no fato da ausência de *frameworks* disponíveis para as Taxonomia Léxicas e o elevado custo computacional para o cálculo das variáveis baseadas em Recuperação da Informação.

De acordo com a revisão bibliográfica realizada neste estudo, não foi possível identificar um padrão na escolha dos grupos de variáveis. Por esse motivo, propomos neste trabalho contemplar os grupos de variáveis mais utilizados na literatura (classes gramaticais, frequência de palavras com N-Gram, medidas de complexidade, palavras-função e emoções), e realizar uma avaliação detalhada sobre a influência destes grupos na tarefa de inferir a idade de usuários a partir de posts.

V. SOLUÇÃO PROPOSTA

Esta seção descreve a metodologia experimental adotada para realização da investigação da influência das variáveis. Inicialmente é descrita a base de dados utilizada. Em seguida, o classificador selecionado é descrito. Por fim, as variáveis que foram criadas para cada grupo são descritas.

A. Base de Dados

A base de dados da competição internacional PAN@CLEF 2013 foi utilizada para investigar a importância dos grupos de variáveis. De acordo com [16], os dados disponíveis em arquivo XML foram selecionados de postagens em blogs cujos autores possuíam idade disponível. O formato dos arquivos é ilustrado na Figura I. Os posts considerados neste estudo foram os escritos em inglês. A base de dados foi dividida em três grupos: treino, avaliação e teste. Cada autor foi aleatoriamente atribuído a um destes três grupos, de modo que todos os seus posts estivessem em apenas um deles. A variável idade foi discretizada em três classes: 10 anos (13-17), 20 anos (23-27) e 30 anos (33-47). A Tabela II descreve as quantidades de posts na língua inglesa para as categorias de idade. Nesta pesquisa, os posts, presentes no conjunto de treino, foram usados para desenvolver os modelos, enquanto que os presentes no conjunto de testes foram usados para verificar o poder de generalização do modelo. Os posts presentes no conjunto de avaliação, entretanto, não foram considerados, uma vez que só eram utilizados, durante a competição, para calibrar os modelos.

```
<author
  lang="lang_code"
  gender="gender_code"
  age_group="age_group">
    <conversations
      count="number_of_conversations_in_file">
      <conversation id="UUID">
        [Original HTML Content of the conversation]
      </conversation>
      <conversation id="UUID">
        [Original HTML Content of the conversation]
      </conversation>

      ....

    </conversations>
</author>
```

Fig. 1. Exemplo de arquivo XML disponibilizado pela competição.

B. Classificador Utilizado

O *Random Forest* foi o classificador selecionado para mensurar a importância das variáveis. Ele é um classificador de aprendizagem combinada (do inglês *Ensemble Learning*) do tipo *Bagging* proposto por [17]. Esse tipo de classificador é uma junção de vários classificadores responsáveis por gerar uma saída de forma individual que será combinada com o objetivo de apresentar uma classificação final [18]. O *Ensemble Learning* combina classificadores que, isoladamente, não apresentam um bom desempenho, mas quando agrupados são capazes de obter um desempenho melhor. No contexto de *Ensemble Learning*, *Random Forest* é um conjunto de árvores de decisão que compõe uma floresta. Cada árvore da floresta será construída independentemente

A justificativa pela escolha do *Random Forest* deve-se ao fato de ele ser apto para bases de dados com alta dimensionalidade e que possuam atributos contínuos e discretos como apontado por [19], [20]. Além de sua adequação a base de dados deste estudo, o *Random Forest* tem um bom desempenho quando comparado com outros classificadores tradicionais, como o k vizinhos mais próximos (KNN) e Redes Neurais Artificiais [5].

C. Grupos de Variáveis Utilizados

O objetivo desta subseção é descrever como cada grupo de palavras foi construído, permitindo assim a replicação e ou melhoria deste experimento por outros pesquisadores.

1) Complexidade: Seis atributos de complexidade foram calculados: número total de caracteres, de palavras, de sentenças. As razões que representam a diversidade do texto são mostradas nas equações: (1), (2) e (3).

$$\frac{\text{Total de Palavras unicas}}{\text{Total de palavras}} \tag{1}$$

$$\frac{\text{Total de caracteres unicas}}{\text{Total de palavras}} \tag{2}$$

$$\frac{\text{Total de palavras}}{\text{Total de sentenças}} \tag{3}$$

Além destes, mais dois atributos ligados à Testes de Leitura foram construídos. Eles indicam o quão fácil um texto pode ser compreendido. São o Teste de Leitura Flesch (Flesch Reading Ease, FRE) e o Teste de Classificação Flesch-Kincaid (Flesch-Kincaid Grading Level, FKLG). O FRE (4) indica a facilidade de leitura de um texto, com valores menores indicando maior complexidade. Já o FKGL (5) indica em que série do sistema educacional americano o texto se adéqua, um valor maior indicando maior complexidade. As equações (4) e (5) exibem as fórmulas de cálculo.

$$FRE = 0,39 * \left(\frac{\text{palavras}}{\text{sentenças}}\right) - 11,8 * \left(\frac{\text{silabas}}{\text{palavras}}\right) - 15,59 \quad (4)$$

$$FKGL = 206.835 - 1.015 * \left(\frac{\text{palavras}}{\text{sentenças}}\right) - 84,6 * \left(\frac{\text{silabas}}{\text{palavras}}\right) \quad (5)$$

2) Palavras-função: Uma lista de palavras-função em inglês foi obtida de *Sequence Publishing* [2], que disponibiliza em seu website um conjunto de dicionários na língua inglesa para propósitos acadêmicos. As palavras-função consideradas neste estudo foram: pronomes, determinadores, conjunções, verbos auxiliares e preposições. A lista utilizada neste estudo foi de 223 palavras-função.

3) Classes gramaticais: As frequências de oito classes gramaticais foram consideradas: verbos, substantivos, adjetivos, advérbios, artigos, pronomes, preposições e conjunções. Para o cálculo deste grupo de variáveis foi utilizado o *Natural Language Toolkit* (NLTK). Ele oferece funcionalidades para a classificação de cada palavra em sua respectiva classe, incluindo um *corpus* de palavras já anteriormente classificadas e um guia para usar o toolkit para este fim [21].

4) Emoções: O estudo utilizou dez variáveis para a identificação de emoções no texto: *positive, negative, joy, surprise, fear, sadness, anger, disgust, trust,* e *anticipation*. Estas variáveis foram construídas a partir do *NRC Emotion Lexicon*. O dicionário NRC é utilizado para detectar a emoção associada a uma palavra [22]. A biblioteca funciona da seguinte forma: caso exista associação entre uma palavra e uma das emoções, o dicionário indica 1, caso contrário, 0. O valor de cada variável associada a respectiva emoção de um documento será a soma dos valores indicados no dicionário para todas as suas palavras.

5) N-Gram: Foram criadas variáveis referentes a 2-Gram e 1-Gram.

VI. Resultados

Foram realizados experimentos para cada grupo de variáveis, conforme descrito na metodologia experimental. A Figura II exibe os resultados obtidos. Os resultados mostram que o grupo de variáveis relacionadas a emoções é o que possui o maior poder discriminatório em relação a tarefa de inferência de idade. A justificativa mais plausível para essa contribuição é que adolescentes e jovens adultos colocam muitas vezes em evidência suas emoções quando estão escrevendo nas redes sociais. No entanto, é importante destacar

[2]https://www.sequencepublishing.com

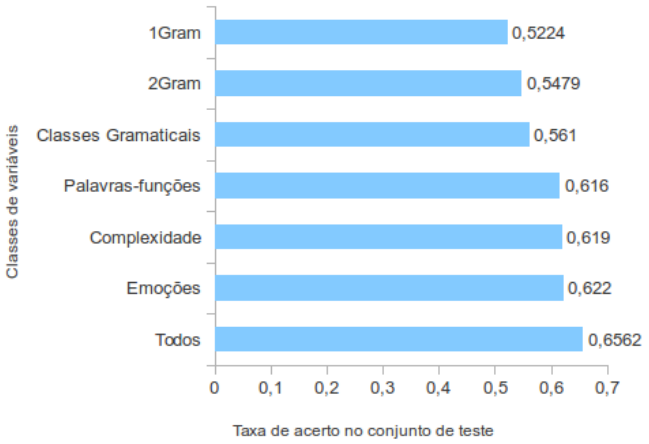

Fig. 2. Resultado por grupo de variáveis.

que outros autores indicaram que variáveis relacionadas a emoções não eram relevantes para este tipo de tarefa, como exemplo, podemos citar [15].

O segundo grupo de varáreis com maior poder discriminatório foi complexidade. A relevância deste grupo de variáveis pode ser explicada porque existe, de uma forma geral, uma correlação direta entre a idade de uma pessoa e o seu nível de escolaridade. Em [15], no entanto, os autores afirmam que este grupo de variáveis obteve um desempenho ruim. As conclusões de [15] sobre a relevância dos grupos de variáveis complexidade e emoções foram prejudicadas pelo fato da natureza das variáveis que foram selecionadas não contribuírem para o classificador que eles utilizaram.

O terceiro grupo de variáveis mais relevante foi palavras-funções e em quarto ficou o grupo de variáreis classes gramaticais. A diferença de desempenho entre estes dois grupos de variáveis pode ser justificada porque o grupo de palavras-funções gera mais variáveis e com isso aumenta a chance de disponibilizar mais informações do que grupo de classes gramaticais.

O grupo de variáveis com pior desempenho foi referente a frequência de palavras 1-Gram e 2-Gram. Este resultado indica que a frequência de termos em texto não é um bom discriminante para inferência de idade. Apesar deste desempenho, não é possível generalizar o resultado obtido para um N maior que 2, uma vez que este estudo não utilizou valores maiores de N. Estudos apontam que a frequência de palavras é relevante para outras tarefas [1]. Por fim, os resultados mostram que a solução com todas as variáveis é a que apresenta o maior poder discriminatório. A explicação para este maior poder preditivo utilizando todas as variáveis é que elas combinadas geram uma informação adicional. Para uma melhor ilustração desta justificativa, vamos tomar como base o clássico problema XOR (o ou exclusivo). O XOR é um problema de classificação binária que possui duas variáveis como entrada e que não é possível criar um classificador que solucione o problema utilizando apenas uma das variáveis, no entanto, quando um classificador não linear recebe as duas entradas ele é capaz de resolver este problema, porque as duas variáveis combinadas geram uma informação adicional ao problema.

VII. Considerações Finais

Este trabalho realizou uma pesquisa sobre a influência dos principais grupos de variáveis na tarefa de inferir a idade de uma pessoa a partir de conteúdo de posts. O estudo foi realizado utilizando uma base de dados de importante competição internacional, considerada um *benchmark* da área. Como metodologia experimental, foi utilizado o classificador *Random Forest* em cinco grupos de variáveis: emoções, complexidade, palavras-funções, classes gramaticais e frequência de palavras. O estudo mostrou que existe diferença de desempenho em um classificador utilizando os diferentes grupos de variáveis.

A principal contribuição deste estudo foi analisar cada grupo de variáveis sob o mesmo classificador e que o mesmo era apropriado para o formato das variáveis, permitindo uma análise sem viés estatístico para identificar diferente relevância entre grupos de variáveis previamente apontados na literatura. Em [15], os autores utilizaram modelos de Regressão Logística para mensurar a importância destes grupos de variáveis. Em um modelo de Regressão Logística, não deve haver multicolinearidade, que ocorre quando as variáveis independentes são correlacionadas uma com as outras [23]. No entanto, as variáveis de um mesmo grupo possuem correlação entre elas. Para ilustrar, vamos analisar o grupo de variáveis Emoções. Uma mesma palavra pode ser associada a mais de uma emoção (variável do grupo). Por exemplo, a palavra *hated* é associada às emoções: *Anger*, *Disgust*, *Negative*, e *Sadness*. O problema relacionado a multicolinearidade não ocorre com o classificador *Random Forest*.

Uma segunda contribuição que pode ser destacada é que a metodologia de construção de variáveis utilizada neste estudo pode ser generalizada para outras tarefas de mineração de texto, como exemplo, podemos citar a tarefa de inferir se um texto de um post escrito em Português é de um Brasileiro ou de um Português. Esta tarefa foi uma subatividade da PAN@CLEF 2017, com os autores deste artigo tendo conseguido a melhor pontuação na classificação de variedades desta língua, empatados com outras duas equipes [24].

Como trabalhos futuros, propomos analisar o desempenho destes grupos de variáveis em uma outra tarefa de mineração de texto como, por exemplo, identificar se uma notícia é hiperpartidária, que significa verificar se ela possui fidelidade cega, preconceituosa, ou irracional a um partido, grupo, causa, ou pessoa [25].

Referências

[1] P. Rosso, F. Rangel, B. Ghanem, and A. Charfi, "ARAP: arabic author profiling project for cyber-security," *Procesamiento del Lenguaje Natural*, vol. 61, pp. 135–138, 2018. [Online]. Available: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5654

[2] P. Rosso, F. Pardo, I. Hernandez-Farias, L. Cagnina, W. Zaghouani, and A. Charfi, "A survey on author profiling, deception, and irony detection for the arabic language," *Language and Linguistics Compass*, vol. 12, no. 4, 2018. [Online]. Available: https://doi.org/10.1111/lnc3.12275

[3] M. Koppel, J. Schler, and S. Argamon, "Computational methods in authorship attribution," *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, no. 1, pp. 9–26, Jan. 2009. [Online]. Available: http://dx.doi.org/10.1002/asi.v60:1

[4] M. A. Hearst, "Untangling text data mining," in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ser. ACL '99. Stroudsburg, PA, USA: Association for Computational Linguistics, 1999, pp. 3–10. [Online]. Available: https://doi.org/10.3115/1034678.1034679

[5] R. F. Oliveira Neto, P. J. L. Adeodato, and A. C. Salgado, "A framework for data transformation in credit behavioral scoring applications based on model driven development," *Expert Syst. Appl.*, vol. 72, pp. 293–305, 2017. [Online]. Available: https://doi.org/10.1016/j.eswa.2016.10.059

[6] M. Meina, K. Brodzinska, B. Celmer, M. Czoków, M. Patera, J. Pezacki, and M. Wilk, "Ensemble-based classification for author profiling using various features," in *Proceedings of Notebook for PAN at CLEF 2013*, ser. CLEF '13. CEUR, 2013, pp. 369–378.

[7] P. Oliveira Lima Junior, L. Gonzaga de Castro Junior, and A. Luiz Zambalde, "Applying textmining to classify news about supply and demand in the coffee market," *IEEE Latin America Transactions*, vol. 14, no. 12, pp. 4768–4774, Dec 2016.

[8] T. Shepherd, "O estatuto da linguística de corpus: metodologia ou área da linguística?" *Matraga - Revista do Programa de Pós-Graduação em Letras da UERJ*, vol. 16, no. 24, 2009.

[9] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler, "Automatically profiling the author of an anonymous text," *Communications of the ACM*, vol. 52, no. 2, pp. 119–123, 2009.

[10] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, ser. Springer Series in Statistics. Springer, 2009.

[11] J. Otterbacher, "Inferring gender of movie reviewers: Exploiting writing style, content and metadata," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, ser. CIKM '10. New York, NY, USA: ACM, 2010, pp. 369–378. [Online]. Available: http://doi.acm.org/10.1145/1871437.1871487

[12] S. Argamon, M. Koppel, J. Fine, and A. R. Shimoni, "Gender, genre, and writing style in formal written texts," *TEXT*, vol. 23, pp. 321–346, 2003.

[13] D. D. Pham, G. B. Tran, and S. B. Pham, "Author profiling for vietnamese blogs," in *Proceedings of the 2009 International Conference on Asian Language Processing*, ser. IALP '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 190–194. [Online]. Available: http://dx.doi.org/10.1109/IALP.2009.47

[14] K. Alsmearat, M. Al-Ayyoub, R. Al-Shalabi, and G. Kanaan, "Author gender identification from arabic text," *Journal of Information Security and Applications*, vol. 35, pp. 85–95, 2017.

[15] E. Weren, A. Kauer, L. Mizusaki, V. Moreira, d. O. Palazzo, and L. Wives, "Examining multiple features for author profiling," *Journal of Information and Data Management*, vol. 5, no. 3, pp. 266–279, 2014.

[16] F. RANGEL and P. ROSSO, "Use of language and author profiling: Identification of gender and age," in *Proceedings of 10th International Workshop on Natural Language Processing and Cognitive Science*, ser. NLPCS '13. Bernade(e Sharp and Michael Zock, 2013, pp. 177–186.

[17] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[18] R. Mendes and R. O. Neto, "The power of ensemble models in fingerprint classification: A case study," *INFOCOMP*, vol. 17, no. 1, pp. 1–10, 2018.

[19] Y. Qi, "Random forest for bioinformatics," in *Ensemble Machine Learning*, C. Zhang and Y. Ma, Eds. Springer US, 2012, pp. 307–323.

[20] H. Finch, "A comparison of methods for group prediction with high dimensional data," *Journal of Modern Applied Statistical Methods*, vol. 13, 2014.

[21] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, 1st ed. O'Reilly Media, Inc., 2009.

[22] S. Mohammad, P. Sobhani, and S. Kiritchenko, "Stance and sentiment in tweets," *Special Section of the ACM Transactions on Internet Technology on Argumentation in Social Media*, vol. 17, no. 3, 2017.

[23] B. Bowerman and R. O Connell, *Linear Statistical Models: An Applied Approach*, ser. Classic Series. Duxbury, 2000.

[24] R. R. Oliveira and R. F. O. Neto, "Using character n-grams and style features for gender and language variety classification," in *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017.*, 2017.

[25] J. Kiesel, M. Mestre, R. Shukla, E. Vincent, D. Corney, P. Adineh, B. Stein, and M. Potthast, "Data for pan at semeval 2019 task 4: Hyperpartisan news detection," nov 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1489920

Rosalvo Neto Possui graduação em Sistemas de Informação pela Faculdade Integrada do Recife (2004), mestrado e doutorado em Ciências da Computação pela Universidade Federal de Pernambuco em 2008 e 2015 respectivamente. Atualmente é professor da Universidade Federal do Vale do São Francisco (UNIVASF). Tem experiência na área de Ciência da Computação, com ênfase em Mineração de Dados, Inteligência Artificial, Banco de Dados e Engenharia de Software, atuando principalmente nos seguintes temas: processo de descoberta de conhecimento em bases de dados, algoritmos de buscas, data warehouse e desenvolvimento dirigido por modelos.

Rodrigo Ribeiro Oliveira Possui graduação em Engenharia da Computação pela Universidade Federal do Vale do São Francisco (UNIVASF) (2018). Atualmente é mestrando em Ciência da Computação na Universidade Estadual de Feira de Santana (UEFS). Tem experiência na área de Ciência da Computação, com ênfase em Inteligência Artificial, Ciência de Dados e Mineração de Dados, atuando principalmente nas seguintes áreas: aprendizado de máquina, text mining e caracterização de autor.

Ana Emília Possui graduação em Ciências da Computação pela Universidade Federal de Pernambuco (2001), mestrado em Ciências da Computação pela Universidade Federal de Pernambuco (2006) e doutorado em Psicologia pela Universidade Federal do Espírito Santo. Atualmente é professora Adjunta II da Universidade Federal do Vale do São Francisco, no colegiado de Engenharia da Computação. Tem experiência na área de Ciência da Computação, com ênfase em Sistemas de Informação, atuando principalmente nos seguintes temas: Engenharia de Software, Software Educativo, Análise Qualitativa da Usabilidade e da Aprendizagem.