UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

EDSON ROBERTO DUARTE WEREN

Atribuição de Perfis de Autoria

Dissertação apresentada como requisito parcial
para a obtenção do grau de
Mestre em Ciência da Computação

Prof. Dra. Viviane Pereira Moreira
Orientador

Prof. Dr. José Palazzo Moreira de Oliveira
Co-orientador

Porto Alegre, novembro de 2014

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Weren, Edson Roberto Duarte

Atribuição de Perfis de Autoria / Edson Roberto Duarte Weren. – Porto Alegre: PPGC da UFRGS, 2014.

50 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2014. Orientador: Viviane Pereira Moreira; Coorientador: José Palazzo Moreira de Oliveira.

1. Armazenamento e Recuperação de Informação. 2. Processamento de Textos e Documentos. I. Moreira, Viviane Pereira. II. de Oliveira, José Palazzo Moreira. III. Título.

“Bom mesmo é ir à luta com determinação, abraçar a vida com paixão, perder com classe e vencer com ousadia,
porque o mundo pertence a quem se atreve. E a vida é muito bela para ser insignificante”

— SIR CHARLES SPENCER CHAPLIN

AGRADECIMENTOS

À Prof^a. Dr^a. Viviane Moreira Pereira (Orientadora), por me guiar em todo o processo, por ir muito além do que se espera de um orientador, deixo aqui toda a minha consideração, respeito e carinho. Seus comentários, sugestões, discussões e revisões, sempre muito relevantes, contribuíram imensamente na condução do trabalho desenvolvido.

Ao Prof. Dr. José Palazzo Moreira de Oliveira (Coorientador), por confiar que eu teria competência para levar este mestrado até o fim, além de compartilhar seu conhecimento e visão estratégica.

À Prof^a. Dr^a. Renata de Matos Galante, pelas ajudas nas questões administrativas do mestrado e atividade didática.

Ao Prof. Dr. Leandro Krug Wives, pelas importantes sugestões e contribuições em coautoria do artigo JIDM relativo a parte desta dissertação.

Aos colegas Anderson Kauer e Lucas Mizusaki, por participarem na avaliação experimental apresentada no artigo JIDM, de grande valia nesta etapa.

Aos membros da banca de defesa desta dissertação, por terem aceitado o convite, pelas sugestões que contribuíram no aprimoramento do texto final e pelo incentivo para o seguimento da pesquisa realizada.

À empresa Eletrobras CGTEE na figura de meus supervisores ao longo destes 24 meses: Annete Picolli, Sergio Santos e Rosângela Machado, por terem me dado a oportunidade e condições para que eu realizasse o mestrado.

Aos meus pais Luis e Venilda, de origem humilde mas com uma visão singular sobre a vida, sempre transmitindo motivação para, em muitas ocasiões, seguir em frente desenvolvendo meu trabalho.

E, por fim, à minha esposa Tatiana, por ter sido companheira, paciente e ouvinte de minhas apreensões e dificuldades, compreendendo e valorizando a dedicação com a qual tive de me entregar a este trabalho.

Author Profiling

ABSTRACT

Authorship analysis aims at classifying texts based on the stylistic choices of their authors. The idea is to discover characteristics of the authors of the texts. This task has a growing importance in forensics, security, and marketing. In this work, we focus on discovering age and gender from blog authors. With this goal in mind, we analyzed a large number of features – ranging from Information Retrieval to Sentiment Analysis. This paper reports on the usefulness of these features. Experiments on a corpus of over 236K blogs show that a classifier using the features explored here have outperformed the state-of-the art. More importantly, the experiments show that the Information Retrieval features proposed in our work are the most discriminative and yield the best class predictions.

Keywords: Information Storage and Retrieval, Document and Text Processing.

RESUMO

A identificação de perfis de autoria visa classificar os textos com base nas escolhas estilísticas de seus autores. A ideia é descobrir as características dos autores dos textos. Esta tarefa tem uma importância crescente em análise forense, segurança e marketing. Neste trabalho, nos concentramos em descobrir a idade e o gênero dos autores de blogs. Com este objetivo em mente, analisamos um grande número de atributos - que variam de recuperação de informação até análise de sentimento. Esta dissertação relata a utilidade desses atributos. Uma avaliação experimental em um corpus com mais de 236K posts de blogs mostrou que um classificador usando os atributos explorados aqui supera o estado-da arte. Mais importante ainda, as experiências mostram que os atributos oriundos de recuperação de informação propostos neste trabalho são os mais discriminativos e produzem as melhores previsões.

Palavras-chave: Armazenamento e Recuperação de Informação, Processamento de Textos e Documentos.

LISTA DE FIGURAS

Figura 3.1: Sistemática de cálculo para as métricas Cosseno e Okapi-BM25 . . . 22
Figura 3.2: Trecho do NRC *emotion lexicon* 25
Figura 3.3: Sistemática de funcionamento do classificador SVM (SEBASTIANI, 2002) 28
Figura 3.4: Exemplo de classificador baseado em regras para a atribuição de perfis de autoria 29
Figura 3.5: Árvore de decisão equivalente à regra DNF 30

Figura 4.1: Exemplo de arquivo de entrada XML. 31
Figura 4.2: Exemplo de matriz de confusão 33
Figura 4.3: Abordagem *wrapper* para seleção de subconjunto de atributos (KOHAVI; JOHN, 1997) 35
Figura 4.4: Busca no espaço de estados para seleção de subconjunto de atributos (KOHAVI; JOHN, 1997) 36
Figura 4.5: Método de validação cruzada para a acurácia estimada considerando 3 execuções (KOHAVI; JOHN, 1997) 36
Figura 4.6: Medidas-F máxima e média, considerando todos/subconjunto de atributos para gênero (a) e idade (b) 37
Figura 4.7: Comparação em termos de acurácia com os resultados oficiais do PAN 40

LISTA DE TABELAS

Tabela 2.1: Comparação sumarizada dos trabalhos relacionados 19

Tabela 3.1: Exemplo de frequência de termos para o vetor de *post* e o vetor de consulta 22

Tabela 4.1: Detalhes dos conjuntos de dados de treinamento e teste do PAN . . . 32
Tabela 4.2: Melhores e piores atributos para idade e gênero de acordo com seu ganho de informação 34
Tabela 4.3: Performance em termos de velocidade dos dez melhores classificadores 38
Tabela 4.4: Resultados da classificação removendo/mantendo cada grupo de atributos 39
Tabela 4.5: Atributos utilizados pelos participantes do PAN2014 41
Tabela 4.6: Distribuição de cada corpus de teste com relação à classe idade por idioma (RANGEL et al., 2014) 42
Tabela 4.7: Resultados médios em termos de acurácia (RANGEL et al., 2014) . . 42
Tabela 4.8: Níveis de significância (RANGEL et al., 2014) 43
Tabela 4.9: Significância das diferenças de acurácia entre os pares de sistemas para identificação de gênero e idade em todo o corpus (RANGEL et al., 2014) 43
Tabela 4.10: Significância das diferenças de acurácia entre os pares de sistemas para identificação de idade em todo o corpus (RANGEL et al., 2014) 43
Tabela 4.11: Significância das diferenças de acurácia entre os pares de sistemas para identificação de gênero em todo o corpus (RANGEL et al., 2014) 44

Tabela 5.1: Classificadores usados nos experimentos 50

LISTA DE ABREVIATURAS E SIGLAS

CLEF	*Conference and Labs of the Evaluation Forum*
CSV	*Categorization Status Value*
DFN	*Disjunctive Normal Form*
FKGL	*Flesch-Kincaid Grade Level*
FRE	*Flesch Reading Ease*
JIDM	*Journal of Information and Data Management*
kNN	*K-Nearest Neighbor*
LIWC	*Linguistic Inquiry and Word Count*
NLTK	*Natural Language Toolkit*
NRC	*National Research Council of Canada*
PAN	*Plagiarism, Authorship, and Social Software Misuse*
RI	Recuperação de Informação
RN	Rede Neural
SBBD	Simpósio Brasileiro de Banco de Dados
SVM	*Support Vector Machine*
WEKA	*Waikato Environment for Knowledge Analysis*

SUMÁRIO

1 INTRODUÇÃO 12
1.1 Motivação 13
1.2 Objetivo 13
1.3 Contribuições 14
1.4 Publicações 14
1.5 Organização da Dissertação 14

2 TRABALHOS RELACIONADOS 16

3 ATRIBUIÇÃO DE PERFIS DE AUTORIA 20
3.1 Atributos 20
3.1.1 Tamanho/Comprimento 20
3.1.2 Recuperação de Informação 20
3.1.3 Legibilidade 24
3.1.4 Análise de Sentimento 24
3.1.5 Corretude 25
3.1.6 Estilo 26
3.2 Classificadores 26
3.2.1 Bayes 26
3.2.2 Funções 27
3.2.3 Lazy 28
3.2.4 Meta 28
3.2.5 Misc 28
3.2.6 Regras 29
3.2.7 Árvores 29

4 EXPERIMENTOS 31
4.1 Materiais e Métodos 31
4.1.1 Conjunto de dados 31
4.1.2 Métricas de Avaliação 32
4.2 Resultados 33
4.2.1 Seleção de Atributos 33
4.2.2 Classificadores 37
4.2.3 Atributos por Grupo 38
4.2.4 Comparação com os Resultados Oficiais do PAN2013 39
4.2.5 Desempenho em Termos de Velocidade de Processamento 39
4.2.6 PAN2014 40

5 CONCLUSÃO 45
REFERÊNCIAS 47
ANEXO I 50

1 INTRODUÇÃO

A identificação de perfis de autoria visa descobrir o máximo de informações possível sobre uma pessoa apenas por meio da análise de textos escritos por ela. A análise pode se concentrar em identificar a idade, sexo, língua materna, nível de educação, ou outras categorias socioeconômicas. Este é um campo fértil, com uma ampla gama de aplicações como: análise forense, marketing e segurança na Internet. Por exemplo, as empresas podem analisar os textos dos produtos avaliados para identificar qual o tipo de cliente gosta ou não gosta de seus produtos, ou a polícia pode identificar o autor de um crime por meio da análise da escrita dos suspeitos.

Essa área vem ganhando maior atenção nos últimos dois anos. Como consequência, em 2013 no PAN *Workshop series on uncovering plagiarism, authorship, and social software misure* [1] foi criado um laboratório de avaliação com foco na identificação de perfis de autoria (GOLLUB et al., 2013); (RANGEL et al., 2013). Neste sentido, os perfis são as classes que precisam ser previstas. Essa tarefa exigia que os participantes construíssem estratégias para identificar o gênero (masculino / feminino) e faixa etária (10s, 20s, 30s) dos autores de um conjunto de *posts* de blogs dados como entrada. A tarefa atraiu muito interesse - mais de 70 grupos se inscreveram e 21 efetivamente participaram, enviando um *software*. Os organizadores do PAN forneceram aos participantes os dados de treinamento (*posts* de blogs para os quais a idade e sexo dos autores eram conhecidos) e, em seguida, avaliaram os *softwares* apresentados com um novo conjunto de dados para os quais a classe é desconhecida. O desempenho dos softwares apresentados é classificado de acordo com a acurácia de suas previsões. O melhor sistema para o corpus de teste no idioma inglês foi apresentado por MEINA et al. (2013), alcançando acurácia de 0,59 na previsão de gênero e acurácia de 0,65 para a idade.

No PAN2013, WEREN; MOREIRA; OLIVEIRA (2013) também participamos, com a introdução de dez atributos baseados em recuperação de informação (RI), até então inédito na literatura, produzidos pelas métricas Cosseno ou Okapi BM25.

RI é um processo complexo que tem o objetivo de produzir uma função de *ranking*, ou seja, uma função que atribui escores a documentos em relação a uma consulta. Neste sentido atributos baseados em RI são produzidos com base em operações (contagem, soma e média) sobre estes escores. A sistemática de produção de atributos baseados em RI pode ser visto na Figura 3.1

Os atributos que são baseados no *ranking* produzido pelo cosseno foram os seguintes:

- 10s_cosseno_contagem,
- 20s_cosseno_contagem,

[1] http://pan.webis.de/

- 30s_cosseno_contagem,
- female_cosseno_contagem,
- male_cosseno_contagem.

Os atributos que são baseados no *ranking* produzido pelo BM25 foram os seguintes:

- 10s_okapi_contagem,
- 20s_okapi_contagem,
- 30s_okapi_contagem,
- female_okapi_contagem,
- male_okapi_contagem,

Além do uso de atributos de RI, foram utilizados dois atributos com base em testes de legibilidade(KINCAID et al., 1975) a saber:

- *Flesch Reading Ease* (FRE),
- *Flesch-Kincaid Grade Level* (FKGL).

Esse foi um estudo preliminar em que foi indexada apenas uma pequena amostra a partir dos dados de treinamento.

Nesta dissertação foram executados experimentos ampliando e avaliando o uso dos atributos propostos baseados em RI, além de atributos baseados em análise de sentimento, Legibilidade, Tamanho/Comprimento, Corretude e Estilo, totalizando 61 atributos.

Neste experimento também foram avaliados dos 55 algoritmos de sete categorias diferentes presentes no WEKA (HALL et al., 2009) , quais são mais adequados à tarefa de identificação de perfis de autoria. Os resultados superam o estado-da-arte, superando também a equipe de maior pontuação no PAN 2013. Além disso, eles mostram que os atributos baseados em RI são os mais discriminativos e produzem os melhores resultados de classificação.

1.1 Motivação

Contribuir para o tema de identificação de perfis de autoria por meio de experimentos com diversos atributos e algoritmos de classificação, avaliando quais atributos são mais úteis para discriminar as características dos autores e que algoritmos de classificação são mais adequados para esta tarefa.

1.2 Objetivo

O principal objetivo deste trabalho é analisar ainda mais a validade dos atributos com base RI, considerando um total de 30 desses atributos. Além disso, ao contrário de trabalhos relacionados, não foi feita análise de etiquetas morfossintáticas, por esta operação ser dispendiosa. Outra diferença é que não usamos os termos do texto como atributos. Tendo como principal vantagem a redução significativa do número de atributos necessários para a previsão de classe.

Analisando as acurácias registradas pelas abordagens apresentadas no PAN2013, é possível notar que a identificação de perfis de autoria é uma tarefa desafiadora e ainda há espaço para melhorias.

Como objetivos específicos destacam-se:

- Testar a utilidade de atributos/grupo de atributos em sete categorias; e
- Testar o desempenho de 55 algoritmos de classificação para a tarefa de identificação de perfis de autoria.

1.3 Contribuições

O nosso estudo é o primeiro a utilizar um sistema de RI para geração de atributos para a classificação de textos, visando a identificação de perfis de autoria. Sendo detectada ainda a possibilidade da exploração das seguintes contribuições:

- Proposta de atributos simples para identificação de perfis de autoria;
- Experimentos com os atributos citados no item anterior, determinando qual atributo e quais conjuntos de atributos são mais significativos para a tarefa de identificação de peris de autoria;
- Experimentos com os 55 algoritmos de classificação presentes na ferramenta WEKA, determinando qual é o mais adequado para a tarefa de identificação de perfis de autoria.

1.4 Publicações

Como produção científica, obtivemos a publicação dos seguintes artigos:

- Publicação de parte do trabalho desenvolvido nesta dissertação como artigo completo no *Journal of Information and Data Management* (JIDM)[2] e apresentada no Simpósio Brasileiro de Banco de Dados (SBBD) 2014 (Weren et. al, 2014). O trabalho intitula-se *"Examining Multiple Features for Author Profiling (WEREN et al., 2014)"* e foi indicado ao prêmio *Best Paper*;
- Participação na campanha PAN 2013, resultando no artigo intitulado *"Using Simple Content Features for the Author Profiling Task"* (WEREN; MOREIRA; OLIVEIRA, 2013);
- Participação na campanha PAN 2014, resultando no artigo intitulado *"Exploring Information Retrieval Features for Author Profiling"* (WEREN; MOREIRA; OLIVEIRA, 2014).

1.5 Organização da Dissertação

Esta dissertação está organizada da seguinte forma: no Capítulo 2, descrevemos os principais trabalhos que orientaram o desenvolvimento da nossa abordagem. No Capítulo 3, descrevemos os métodos propostos para a identificação de perfis de autoria, os atributos

[2] http://seer.lcc.ufmg.br/index.php/jidm

para a identificação de perfis e os classificadores utilizados. No Capítulo 4, apresentamos os experimentos realizados, os resultados alcançados e a comparação com os trabalhos *baseline*. No Capítulo 5 serão discutidas algumas questões relevantes relativas à avaliação experimental. Por fim, no Capítulo 6 apresentamos as conclusões e possibilidades para trabalhos futuros.

2 TRABALHOS RELACIONADOS

Neste capítulo, serão apresentados os principais trabalhos relacionados à abordagem proposta.

KOPPEL; ARGAMON; SHIMONI (2003) foram os pioneiros em identificação de perfis de autoria. Eles mostraram que é possível encontrar padrões em estilos de escrita que podem ser usados para classificar o autor em um grupo demográfico específico. Nesse estudo foi utilizado um grupo de *function words* (ou seja, palavras que não têm significado, mas que servem para expressar relações gramaticais com outras palavras - por exemplo, preposições, pronomes, verbos auxiliares) e análise de etiquetas morfossintáticas, a fim de identificar o gênero do autor e se o texto era de ficção ou não-ficção. Eles utilizaram um total de 1.081 atributos, resultando em uma acurácia superior a 0,80 em 920 documentos do *British National Corpus*. Em um trabalho posterior, ARGAMON et al. (2009) definiram dois grupos básicos de atributos que podem ser usados para a Atribuição de Perfis de Autoria: atributos dependentes de conteúdo (ou seja, certos assuntos, palavras-chave e frases-chave que são usados principalmente por um dos grupos); e os atributos baseados em estilo, que são independentes de conteúdo (por exemplo, comprimento médio de palavras). Experimentos com 19K *posts* de blogs usando o algoritmo de aprendizado *Bayesian Multinomial Regression* atingiram acurácia aproximada de 0,76. Estes autores observaram que determinantes e preposições são marcadores de redação masculina, enquanto que os pronomes são marcadores de escrita feminina. Contrações sem apóstrofos foram definidas como marcadores discriminativos de escrita mais jovem, enquanto que as preposições são mais fortes na escrita de pessoas mais maduras.

MUKHERJEE; LIU (2010) identificaram padrões em etiquetas morfossintáticas e propuseram um método de seleção de atributo. Experimentos mostraram que esse método de seleção pode apresentar até 10% de melhoria e a análise de etiquetas morfossintáticas produzem ganho de cerca de 6%.

OTTERBACHER (2010) concentraram-se também na identificação de gênero, mas em outro conjunto de dados: comentários de cinema. Esse estudo concentrou-se em avaliar as diferenças entre o tipo de escrita para o gênero masculino e feminino, com relação ao estilo, conteúdo e metadados (ou seja, data da postagem, avaliação, número de contribuições). Esses atributos foram utilizados em um classificador baseado em regressão logística. Os melhores resultados foram obtidos quando foram utilizados todos os tipos de atributos. Curiosamente, eles descobriram que o atributo de metadados "avaliação" rendeu bons resultados e que comentários escritos por homens costumam ser mais úteis.

Ainda sobre a previsão de gênero, SARAWGI; GAJULAPALLI; CHOI (2011) realizaram experimentos em blogs e artigos científicos. Esses autores utilizaram gramáticas probabilísticas livres de contexto (RAGHAVAN; KOVASHKA; MOONEY, 2010) para capturar regularidades sintáticas além de padrões superficiais baseados em n-gramas. Eles

constataram que modelos a nível de caractere apresentam melhor desempenho do que os modelos a nível de palavra. Como esperado, a previsão de gênero em artigos científicos foi mais difícil do que em blogs. No entanto, os resultados mostram uma acurácia de 61%, o que é bem melhor do que uma previsão aleatória (isto é, 50%). Isso demonstra que gênero pode ser identificado até mesmo na escrita formal.

NGUYEN; SMITH; ROSÉ (2011) focaram-se na previsão de idade em corpora diferentes: blogs, ligações telefônicas transcritas e posts de fóruns online. Os autores utilizaram o gênero como um atributo combinado com unigramas, análise de etiquetas morfossintáticas e classes de palavras obtidas por meio do *Linguistic Inquiry and Word Count* (LIWC) (PENNEBAKER; FRANCIS; BOOTH, 2001). Experiências usando regressão linear mostraram que a combinação de atributos produz melhores resultados. O gênero como atributo mostrou-se muito útil para discriminar a idade de autores mais jovens.

PEERSMAN; DAELEMANS; VAN VAERENBERGH (2011) estudaram a previsão de idade e gênero em redes sociais. O aspecto-chave aqui é o curto tamanho do texto (12 palavras, em média). Os atributos utilizados incluem *emoticons*, sequências de palavras (unigramas, bigramas e trigramas) e caracteres (bigramas, trigramas e tetragramas). Esses atributos foram submetidos a um classificador *Support Vector Machine* (SVM). Atributos de sequências de palavras e *emoticons* mostraram-se mais úteis em relação aos demais atributos. Os melhores resultados foram obtidos quando o conjunto de dados de treinamento era equilibrado.

No ano passado, durante o PAN de 2013, foi criado um ponto comum de referência para avaliar abordagens para identificação de perfis de autoria. Os 21 participantes que apresentaram software relataram suas experiências em trabalhos de *notebooks* que foram resumidos por RANGEL et al. (2013). Os três melhores sistemas classificados por suas acurácias foram MEINA et al. (2013), LÓPEZ-MONROY et al. (2013) e MECHTI; JAOUA; BELGUITH (2013).

MEINA et al. (2013) utilizaram uma ampla gama de atributos, incluindo atributos estruturais (número de frases, palavras, parágrafos), análise de etiquetas morfossintáticas, legibilidade, palavras de emoção, *emoticons*, e tópicos estatísticos derivados de análise semântica latente. Nesse trabalho foram empregados 311 atributos para a idade e 476 para o gênero, que foram utilizados por um classificador *Random Forest*. Com relação ao emprego de alguma técnica de pré-processamento, esta foi utilizada para identificar *posts* com spam e descartá-los do conjunto de dados. Apesar de apresentar os melhores resultados em termos de acurácia, o sistema construído foi um dos mais lentos, levando 4,44 dias para processar os dados de teste (que não incluíam os dados de treinamento). Apresentou a melhor acurácia para gênero e a segunda melhor para a idade.

LÓPEZ-MONROY et al. (2013) propuseram uma abordagem diferente para a representação de documentos, em contraste com abordagens que representam documentos como vetores, utilizando cada termo como um atributo, denominado Atributo de Segunda Ordem. A ideia principal é calcular como cada termo se relaciona com cada perfil. Os perfis são as classes que desejamos prever, a saber: 10s_female, 10s_male, 20s_female, 20s_male, 30s_female e 30s_male. A relação entre um termo e um perfil é baseada na frequência do termo. Uma vez que os vetores de termos são calculados, a relação entre os vetores e perfis de documentos é calculada. Esta abordagem realiza a classificação em seis classes (gênero/idade). Um classificador *LibLinear* (FAN et al., 2008) foi utilizado com os 50 mil termos mais frequentes como atributos. O desempenho dessa abordagem em termos de velocidade do sistema foi interessante, levando 38 minutos para classificar os dados de teste. Esse sistema apresentou a melhor acurácia para idade e a terceira

melhor para o gênero.

MECHTI; JAOUA; BELGUITH (2013) calcularam os 200 termos mais frequentes por perfil. Esses termos foram agrupados em classes, tais como: determinantes, preposições, pronomes, palavras relacionadas ao amor, palavras usadas por adolescentes, etc, totalizando 25 classes para os dados de treinamento em inglês. Os atributos foram usados para treinar um classificador de árvore de decisão (J48). Esse foi o sistema mais lento, levando aproximadamente 11 dias para classificar os dados de teste. Esse sistema apresentou a segunda melhor acurácia para gênero, no entanto a previsão para idade não apresentou acurácia semelhante.

Também participamos nesta competição com dez atributos baseados em RI e dois atributos baseados em testes de legibilidade (WEREN; MOREIRA; OLIVEIRA, 2013). Obtendo o 15^o lugar para gênero e idade na língua inglesa e os 13^o e 14^o lugares respectivamente para gênero e idade na língua espanhola. Em relação ao tempo de processamento nossa abordagem demandou 3 horas e 25 minutos para classificar o corpus de teste.

Esse foi um estudo preliminar, no qual indexamos apenas uma pequena amostra a partir dos dados de treinamento.

Analisando a Tabela 2.1, conclui-se que é difícil estabelecer uma correlação entre os atributos utilizados nas diferentes abordagens e os resultados obtidos, devido, principalmente, à quantidade de atributos comuns e à quantidade heterogênea de corpora utilizados.

Atributos estilísticos e de conteúdo foram utilizados pela grande maioria das abordagens e os valores de exatidão obtidos mostram resultados em diferentes amplitudes. Um atributo comumente utilizado é o derivado de análise de etiquetas morfossintáticas, que foi usado por quatro trabalhos relacionados diferentes. Esse atributo parece melhorar o desempenho da predição de classe.

Com relação ao emprego de algum tipo de pré-processamento que também pode ser visto na Tabela 2.1, é interessante que técnicas de seleção de subconjunto de atributos foram estudadas em três trabalhos relacionados e os resultados demonstram que esta técnica tende a melhorar os resultados.

Neste trabalho, ao contrário de trabalhos relacionados, não fazemos uso de análise de etiquetas morfossintáticas por esta ser uma operação dispendiosa. Outra diferença é que não usamos os termos do texto como atributos.

Diante do exposto, existe a oportunidade de analisarmos ainda mais a validade dos atributos baseados em RI, considerando um total de 30 desses atributos.

Tabela 2.1: Comparação sumarizada dos trabalhos relacionados

	Autores										
Técnicas de pré-processamento utilizadas pelos autores dos trabalhos relacionados	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Remoção de HTMLs			X						X		X
Seleção de Subconjunto			X	X	X						
Discriminação entre humano, spam e chatbots									X		
Atributos de estilo utilizados pelos autores dos trabalhos relacionados	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Frequência de Sinais de Pontuação									X		
Comprimento médio das palavras		X									
Uso de maiúsculas									X		
Citações									X		
Atributos de conteúdo utilizados pelos autores dos trabalhos relacionados	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Análise Semântica Latente									X	X	
Frases-chaves		X									
Palavras-chaves		X									
Bag-of-words									X	X	
TF-IDF									X	X	
Palavras baseadas em dicionário									X	X	
Palavras baseadas em tema									X	X	
Palavras baseadas em entropia									X	X	
Palavras denotando emoção									X		
Atributos derivados de outras abordagens utilizados pelos autores dos trabalhos relacionados	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Uso de motor de busca, considerando o texto como uma consulta											X
Gramáticas Probabilísticas Livres de Contexto							X				
Metadados (Data da Postagem, Avaliação, Número de contribuições e etc)			X								
Function Words	X										
Etiquetas morfossintáticas	X			X		X			X		
Atributos baseados em HTML									X		
Atributos baseados em Emoticons					X			X			
Atributos de Legibilidade									X		X
Uso de modelos n-gramas					X	X			X		
Representação de segunda ordem utilizado com base nas relações entre os documentos e perfis								X			

Legenda Autores:
1-(KOPPEL et al., 2003) **2**-(ARGAMON et al., 2009) **3**-(OTTERBACHER, 2010) **4**-(MUKHERJEE; LIU, 2010) **5**-(PEERSMAN; DAELEMANS; VAN VAERENBERGH, 2011) **6**-(NGUYEN; SMITH; ROSÉ, 2011) **7**-(SARAWGI; GAJULAPALLI; CHOI, 2011) **8**-(LÓPEZ-MONROY et al., 2013) **9**-(MEINA et al., 2013) **10**-(MECHTI; JAOUA; BELGUITH, 2013) **11**-(WEREN; MOREIRA; OLIVEIRA, 2013)

3 ATRIBUIÇÃO DE PERFIS DE AUTORIA

Neste capítulo serão detalhados os atributos usados em nossa investigação (Seção 3.1), bem como a abordagem de classificação que adotamos (Seção 3.2).

3.1 Atributos

Esta seção descreve os atributos de tamanho/comprimento (Subseção 3.1.1), os atributos baseados em RI (Subseção 3.1.2), os atributos de legibilidade (Subseção 3.1.3), os atributos baseados em análise de sentimento (Subseção 3.1.4), os atributos de corretude (Subseção 3.1.5) e os atributos baseados em estilo (Subseção 3.1.6).

A coleção de *posts* de cada autor foi representada por um conjunto de 61 características (ou atributos), que foram divididos em seis grupos. Em seguida, serão explicados cada um desses grupos.

3.1.1 Tamanho/Comprimento

São atributos simples, que calculam o comprimento absoluto do texto do *post*. Sendo o uso desta categoria de atributos citado na literatura técnica por ARGAMON et al. (2009).

- Número de caracteres;
- Número de palavras;
- Número de sentenças.

3.1.2 Recuperação de Informação

Este grupo de atributos criados por ocasião da pesquisa para esta dissertação e inéditos na literatura representa a nossa hipótese de que os autores do mesmo grupo de gênero ou idade tendem a usar termos semelhantes e que a distribuição desses termos seria diferente entre os gêneros e faixas etárias. O conjunto completo de *posts* (sem *stemming* ou remoção de *stopword*) é indexado por um Sistema de RI. Em seguida, o *post* (por exemplo, blog) que queremos classificar é usado como uma consulta e os k *posts* mais semelhantes são recuperados. Neste sentido, a ideia é que os *posts* que serão recuperados (ou seja, os mais parecidos com a consulta) serão os do mesmo gênero e faixa etária. A classificação é dada pelas métricas Cosseno ou Okapi BM25, como explicado abaixo. Essas métricas foram escolhidas por serem as mais utilizadas por sistemas de RI para gerar o *ranking* de documentos em resposta às consultas. O cosseno é usado no modelo vetorial e o Okapi BM25 é um modelo probabilístico bem estabelecido (MANNING; RAGHAVAN; SCHüTZE, 2008). Foram utilizados 30 atributos baseados

em RI.

-Cosseno

Os atributos que são baseados no *ranking* produzido pelo cosseno são os seguintes:

- 10s_cosseno_contagem,
- 20s_cosseno_contagem,
- 30s_cosseno_contagem,
- female_cosseno_contagem,
- male_cosseno_contagem,
- 10s_cosseno_soma,
- 20s_cosseno_soma,
- 30s_cosseno_soma,
- female_cosseno_soma,
- male_cosseno_soma,
- 10s_cosseno_média,
- 20s_cosseno_média,
- 30s_cosseno_média,
- female_cosseno_média, e
- male_cosseno_média.

Esses atributos são calculados como uma função de agregação sobre os $top\text{-}k$ resultados para cada grupo de idade/gênero obtido em resposta a uma consulta composta pelas palavras-chave do *post*. Testamos três tipos de funções de agregação: contagem, soma e média. Para este conjunto de atributos, consultas e *posts* foram comparados utilizando o cosseno (Eq. 3.1). Por exemplo, se fossem recuperados 10 *posts* em resposta a uma consulta composta por palavras-chave e 5 dos *posts* recuperados estivessem na faixa etária de 30s, o valor para 30s_cosseno_média seria a média dos escores dos *posts* recuperados que se enquadrassem na categoria 30s. Essa sistemática de cálculo pode ser vista na Figura 3.1.

Figura 3.1: Sistemática de cálculo para as métricas Cosseno e Okapi-BM25

A métrica cosseno é dada como:

$$COSSENO = (c,s)\frac{\vec{c}.\vec{q}}{|\vec{c}||\vec{q}|} \tag{3.1}$$

Onde $\vec{c}$ e $\vec{q}$ são os vetores para o *post* e a consulta, respectivamente. Os vetores são compostos pelos pesos $tf_{i,c} \times idf_i$ onde $tf_{i,c}$ é a frequência do termo i no blog c, e idf_i, = $log\frac{N}{n(i)}$ onde N é o total do número de *posts* na coleção, e $n(i)$ é o número de *posts* que contêm i.

Exemplificando:

Na Tabela 3.1 é exibido um exemplo de um vetor de *post* e um vetor de consulta, e a respectiva frequência de cada termo em cada vetor.

Na sequência é mostrada a sistemática de cálculo da métrica cosseno. Quanto mais o resultado se aproximar de 1 mais o *post* e a consuta são similares.

Tabela 3.1: Exemplo de frequência de termos para o vetor de *post* e o vetor de consulta

TERMO	**C**	**Q**
Computação	3	1
RI	2	0
Atributos	0	0
UFRGS	5	0
PPGC	0	0
Cosseno	0	0
Dissertação	2	1
Classificadores	0	0
Legibilidade	0	2

$$\vec{c}.\vec{q} = 3*1+2*0+0*0+5*0+0*0+0*0+0*0+2*1+0*0+0*2 = 5$$

$$|\vec{c}| = (3^2+2^2+0^2+5^2+0^2+0^2+0^2+2^2+0^2+0^2)^{0,5} = (42)^{0,5} = 6,481$$

$$|\vec{q}| = (1^2+0^2+0^2+0^2+0^2+0^2+0^2+1^2+0^2+2^2)^{0,5} = (6)^{0,5} = 2,245$$

$COS(c,q) = 0,344$

-Okapi BM25

Os atributos que são baseados no *ranking* produzido pelo BM25 são os seguintes:

- 10s_okapi_contagem,
- 20s_okapi_contagem,
- 30s_okapi_contagem,
- female_okapi_contagem,
- male_okapi_contagem,
- 10s_okapi_soma,
- 20s_okapi_soma,
- 30s_okapi_soma,
- female_okapi_soma,
- male_okapi_soma,
- 10s_okapi_média,
- 20s_okapi_média,
- 30s_okapi_média,
- female_okapi_média, e
- male_okapi_média.

Semelhante ao conjunto de atributos anterior, esses atributos calculam uma função de agregação (média, soma e contagem) sobre os top-k resultados obtidos de cada grupo gênero/idade em resposta a uma consulta composta pelas palavras-chave no *post*. Para este conjunto de atributos, consultas e *posts* foram comparados com a métrica Okapi BM25 (Eq. 3.2), como segue:

$$BM25(c,q) = \sum_{i=1}^{n} IDF_i \frac{tf_{i,c} \cdot (k_1 + 1)}{tf_{i,c} + k_1(1 - b + b\frac{|D|}{avgdl})} \tag{3.2}$$

Onde $tf_{i,c}$ e idf_i são como na Eq. 3.1, $|D|$ é o comprimento (em palavras) dos *posts*, $avgdl$ é o comprimento médio dos *posts* na coleção, k_1 e b são parâmetros que selecionam a importância da presença de cada termo da consulta e o comprimento do *post*. Nos experimentos realizados, os valores para k_1 e b receberam, respectivamente, os valores 1,2 e 0,75. Estes são os valores padrão utilizados pelo Zettair[1].

[1] http://www.seg.rmit.edu.au/zettair/

3.1.3 Legibilidade

Testes de legibilidade indicam a dificuldade de compreensão de um texto (*post*). Sendo o uso desta categoria de atributos citado na literatura técnica por MEINA et al. (2013).

-Testes de legibilidade Flesch-Kincaid

Nós empregamos dois testes que indicam a dificuldade de compreensão de um *post*: *Flesch Reading Ease* (FRE) e *Flesch-Kincaid Grade Level* (FKGL) (KINCAID et al., 1975). Eles são dados pelas equações. 3.3 e 3.4.

Pontuações mais altas para FRE indicam um material que é mais fácil de ler. Por exemplo, um *post* com uma pontuação FRE entre 90 e 100 poderia ser facilmente lido por um adolescente de 11 anos, enquanto que *posts* com pontuação abaixo de 30 seriam melhor compreendidos por alunos de graduação. Pontuações FKGL indicam o nível de ensino. Um FKGL de 7 indica que o *post* é compreensível por um estudante da $7^{\underline{a}}$ série. Assim, quanto maior a pontuação FKGL, maior é o número de anos de formação necessários para compreender o *post*. A ideia de usar esses índices é ajudar a distinguir a idade do autor. Autores mais jovens tendem a usar palavras mais curtas e, portanto, teriam uma FKGL menor e um FRE alto.

$$FRE = 206.835 - 1.015\left(\frac{\#Palavras}{\#Sentenças}\right) - 84.6\left(\frac{\#Silabas}{\#Palavras}\right) \tag{3.3}$$

$$FKGL = 0.39\left(\frac{\#Palavras}{\#Sentenças}\right) + 11.8\left(\frac{\#Silabas}{\#Palavras}\right) - 15.59 \tag{3.4}$$

3.1.4 Análise de Sentimento

Estes atributos foram extraídos com base no *National Research Council of Canada* (NRC) *Emotion Lexicon* (MOHAMMAD; KIRITCHENKO; ZHU, 2013), que atribui pesos aos termos para refletir o tipo de emoção que eles transmitem. A Figura 3.2 mostra um trecho do NRC *Emotion lexicon*. Por exemplo, a palavra eficiente é considerada uma palavra positiva associada com a emoção confiança. As palavras do *post* foram lematizadas para que elas correspondam à forma canônica em que as palavras são encontradas no dicionário.

Esse processo foi realizado com o auxílio do *Natural Language Toolkit* (NLTK) [2]. Sendo o uso desta categoria de atributos citado na literatura técnica por MEINA et al. (2013).

-NRC *emotions*

Dez dos atributos usados vêm do NRC, a saber:

- *positive*,
- *negative*,
- *joy*,

[2] http://www.nltk.org/

- *surprise*,
- *fear*,
- *sadness*,
- *anger*,
- *disgust*,
- *trust*,
- *anticipation*.

Para calcular os atributos baseados em análise de sentimento, simplesmente buscou-se cada palavra do *post* no NRC *emotion lexicon*. Se o *post* tem uma ou mais emoções associadas, as pontuações são somadas.

efficacy	positive:1	negative:0	anger:0	anticipation:0	disgust:0	fear:0	joy:0	sadness:0	surprise:0	trust:0
efficiency	positive:1	negative:0	anger:0	anticipation:0	disgust:0	fear:0	joy:0	sadness:0	surprise:0	trust:0
efficient	positive:1	negative:0	anger:0	anticipation:1	disgust:0	fear:0	joy:0	sadness:0	surprise:0	trust:1
effigy	positive:0	negative:0	anger:1	anticipation:0	disgust:0	fear:0	joy:0	sadness:0	surprise:0	trust:0
effort	positive:1	negative:0	anger:0	anticipation:0	disgust:0	fear:0	joy:0	sadness:0	surprise:0	trust:0

Figura 3.2: Trecho do NRC *emotion lexicon*

3.1.5 Corretude

Este grupo de atributos visa captar a correção do *post*. Sendo o uso desta categoria de atributos citado na literatura técnica por MEINA et al. (2013) e MECHTI; JAOUA; BELGUITH (2013).

- **Palavras no dicionário**: Relação entre as palavras do *post* encontrado no dicionário OpenOffice US[3] e o número total de palavras no *post*. (Eq. 3.5) como segue:

$$Palavras\ no\ Dicionário = \frac{Número\ de\ Palavras\ Reconhecidas\ no\ Dicionário}{Total\ de\ Palavras} \tag{3.5}$$

- **Limpeza**: Relação entre o número de caracteres no *post* pré-processado e o número de caracteres no *post* bruto. O pré-processamento é detalhado na Seção 4.1. A ideia é avaliar como o *post* original é "limpo". (Eq. 3.6) como segue:

$$Limpeza = \frac{Total\ de\ Caracteres\ do\ post\ Pré\text{-}Processado}{Total\ de\ Caracteres\ do\ post\ Original} \tag{3.6}$$

- **Vogais Repetidas**: Em alguns casos, os autores usam palavras com vogais repetidas para dar ênfase. Por exemplo, *"I am soo tired"*. Este grupo de atributos conta o número de vogais repetidas (a, e, i, o e u) em sequência dentro de uma palavra.

- **Pontuação Repetidas**: Este atributo calcula o número de marcas de pontuação repetidas (ou seja, vírgulas, pontos e vírgulas, pontos finais, pontos de interrogação e pontos de exclamação) em sequência no *post*.

[3] http://extensions.openoffice.org/en/project/english-dictionaries-apache-openoffice

3.1.6 Estilo

Este grupo de atributos visa captar o modo de como cada grupo de autores se expressa em um *post*. Sendo o uso desta categoria de atributos citado na literatura técnica por MEINA et al. (2013)

- **Tags HTML**: Esse atributo consiste em contar o número de *tags* HTML que indicam quebras de linha <br_html>, imagens <img_html> e links <href_html>.

- **Diversidade**: Esse atributo é calculado como a relação entre as palavras distintas no *post* e o número total de palavras no *post*. (Eq. 3.7) como segue:

$$Diversidade = \frac{Número\ de\ Palavras\ Distintas}{Total\ de\ Palavras} \tag{3.7}$$

3.2 Classificadores

Os atributos descritos na Seção 3.1 são usados para treinar algoritmos de aprendizado de máquina supervisionado. Como as classes das instâncias de treinamento são conhecidas o uso de algoritmos de classificação é justificado. O modelo treinado com as instâncias do corpus de treinamento é então usado para classificar novas instâncias (por exemplo, os dados de teste).

Existem vários algoritmos de aprendizado de máquina disponíveis, e avaliações empíricas (CARUANA; NICULESCU-MIZIL, 2006) mostraram que não há um classificador universalmente melhor - mesmo os melhores podem apresentar baixo desempenho para alguns tipos de problemas. Assim, no presente trabalho, avaliou-se uma série de algoritmos com o objetivo de selecionar os que apresentam melhor desempenho para a tarefa de identificação de perfis de autoria.

Depois que os atributos foram computados, realizou-se o treinamento dos classificadores. Weka(WITTEN; FRANK, 2005) foi usado para construir os modelos de classificação. Ele fornece uma série de algoritmos de aprendizado de máquina, divididos em sete grupos, os quais foram sumarizados em um levantamento feito por SEBASTIANI (2002). Deste levantamento, foi possível obter mais informações sobre o funcionamento destes classificadores, os quais serão descritos nas próximas subseções.

3.2.1 Bayes

Contendo classificadores Bayesianos, por exemplo, *NaiveBayes*.

Classificadores Bayesianos são classificadores probabilísticos onde o *Categorization Status Value* $(CSV)_i(d_j)$ em termos de $P(classe|\overrightarrow{d}_j)$ (probabilidade que um documento representado por um vetor $\overrightarrow{d}_j = \langle w_{1j}, ..., w_{|\tau|j} \rangle$ de termos pertença à *classe*) é computado pela aplicação do teorema de Bayes, dado por:

$$P(classe|\overrightarrow{d}_j) = \frac{P(classe)P(\overrightarrow{d}_j|classe)}{P(\overrightarrow{d}_j)} \tag{3.8}$$

Na equação 3.8, o evento espaço é o espaço de documentos, então $P(\overrightarrow{d}_j)$ é a probabilidade que um documento escolhido aleatoriamente tenha o vetor $\overrightarrow{d}_j$ como sua representação, e $P(classe)$ é a probabilidade que um documento escolhido aleatoriamente pertença à *classe*.

A estimativa de $P(\overrightarrow{d}_j|classe)$ na equação 3.8 é problemática, já que o número de possíveis vetores $\overrightarrow{d}_j$ é muito alto. Para atenuar este problema, é comum supor que duas coordenadas no vetor de documentos são variáveis aleatórias, estatisticamente independentes uma da outra. Essa suposição de independência é codificada pela equação:

$$(\overrightarrow{d}_j|classe) = \Pi_{k=1}^{|\tau|} P(w_{kj}|classe) \tag{3.9}$$

Classificadores probabilísticos que usam essa suposição são chamados de classificadores *Naive Bayes*.

3.2.2 Funções

Incluindo *Support Vetor Machines* (SVM), algoritmos de regressão, redes neurais.
A seguir serão descritos os tipos de algoritmos pertencentes a essa categoria:

- Algoritmos de regressão: Regressão denota a aproximação de uma função $\breve{\Phi}$ real por meio de uma função Φ que se ajusta aos dados de treinamento.

- Classificadores Lineares: Um classificador linear para a categoria *classe* é um vetor $\overrightarrow{classe} = \langle w_{1i}, ..., w_{|\tau|i} \rangle$ pertencente ao mesmo espaço $|\tau|$-dimensional, em que os documentos também são representados, e de tal modo que $CSV_i(d_j)$ corresponda ao produto escalar de $\sum_{k=1}^{|\tau|} w_{ki} w_{kj}$ do $\overrightarrow{d}_j$ e $\overrightarrow{classe}$.

  Um simples método linear é o algoritmo *Perceptron*. Neste algoritmo, o classificador para *classe* é primeiro inicializado, definindo todos os pesos w_{ki} para o mesmo valor positivo. Quando um exemplo de treinamento d_j é examinado, o classificador construído até o momento realiza a classificação. Se o resultado da classificação é correto, nada é alterado; por outro lado, se a classificação é incorreta, os pesos do classificador são modificados. Se d_j é um exemplo positivo, os pesos w_{ki} de "termos ativos" são "promovidos" pelo incremento para um valor fixo $\alpha > 0$, enquanto que se d_j for um exemplo negativo então os mesmos pesos são "rebaixados" . Note que quando o classificador atingiu um nível razoável de eficácia, considerando ainda o fato do peso w_{ki} ser muito baixo, significa que t_k contribui até o momento negativamente para o processo de classificação. Este, portanto, pode ser descartado da representação.

- Classificadores de Redes Neurais: Uma Rede Neural (RN) é uma rede de unidades onde as unidades de entrada representam os termos, a(s) unidade (s) de saída representam a categoria ou categorias de interesse, e os pesos nas extremidades que conectam as unidades representam as relações de dependência. Para classificar um documento de teste d_j, estes pesos de termos w_{kj} são carregados entrada; a ativação dessas unidades é propagada através da rede, e o valor de saída da unidade determina a decisão de categorização. Um modo típico de treinamento de RNs é a retro propagação, através do qual os pesos dos termos dos documentos de treinamento são carregados nas unidades de entrada. Se ocorrer um erro de classificação, o erro é "retro propagado" , de modo a alterar os parâmetros da rede e eliminar ou minimizar o erro.

- Classificadores SVM: Em termos geométricos, pode ser visto como uma tentativa para encontrar, entre todas as superfícies $\sigma_1, \sigma_2, ...$ no espaço $|\tau|$-dimensional, a separação entre os exemplos de treinamentos positivos e negativos, o σ_i que separa

os positivos dos negativos pela margem mais ampla o possível. Essa ideia é melhor entendida no caso em que os aspectos positivos e negativos são linearmente separáveis, caso em que as superfícies de decisão são $(|\tau| - 1)$-hiperplanos. No caso de duas dimensões da figura 3.3, várias linhas podem ser escolhidas como as superfícies de decisão. O método SVM escolhe o elemento médio a partir do "Maior" conjunto de linhas paralelas, isto é, a partir do conjunto em que a distância máxima entre os dois elementos do conjunto é mais elevada. Vale ressaltar que essa "melhor" superfície de decisão é determinada por apenas um pequeno conjunto de exemplos de treinamento, chamado de vetores de suporte.

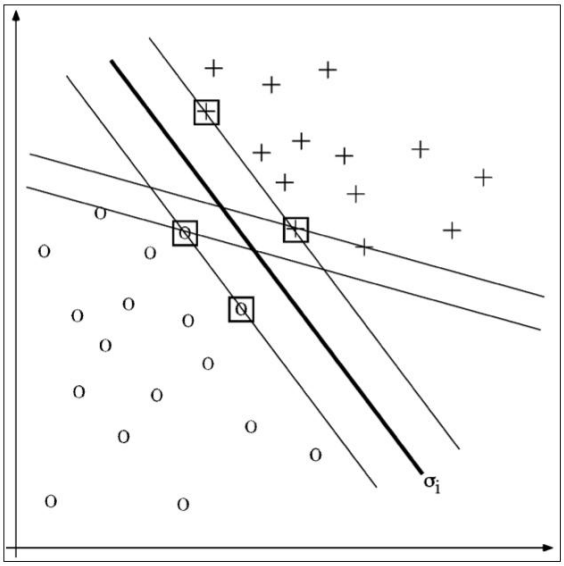


Figura 3.3: Sistemática de funcionamento do classificador SVM (SEBASTIANI, 2002)

3.2.3 Lazy

A aprendizagem é realizada no momento de predição, por exemplo, *K-Nearest Neighbor* (k-NN).

Também chamados de classificadores baseados em exemplos, estes não constroem uma representação explícita, declarativa da categoria c_i, e sim consideram os rótulos de categoria anexados aos documentos de treinamento que são semelhantes no documento de teste.

3.2.4 Meta

Meta-classificadores usam como base um ou mais classificadores como entrada, por exemplo, *boosting*, *bagging* ou *stacking*.

Os Meta-classificadores baseiam-se na ideia de que, dada uma tarefa que exige conhecimento especializado para ser realizada, k especialistas podem ser melhores do que um se seus julgamentos individuais forem combinados apropriadamente. A ideia é aplicar k classificadores diferentes $\Phi_1, ..., \Phi_k$ para a mesma tarefa, decidindo se $d_j \in c_i$ e em seguida combinar seus resultados de forma adequada. O classificador ensemble é caracterizado por (i) a escolha de k classificadores, e (ii) a escolha de uma função de combinação.

3.2.5 Misc

Vários classificadores que não se encaixam em nenhuma outra categoria.

3.2.6 Regras

Classificadores baseados em regras, por exemplo, *ZeroR*.

Nos classificadores baseados em regras, um classificador para a categoria *classe* construído por um método de aprendizagem de indução de regra consiste de uma regra *Disjunctive Normal Form* (DNF), que é uma regra condicional com uma premissa em forma normal disjuntiva, do tipo ilustrado na Figura 3.4. Os literais na premissa denotam a presença ou ausência de uma palavra-chave no documento de teste d_j, enquanto que a cláusula principal denota a decisão de classificar d_j sobre a *classe*. Regras DNF são similares a árvores de decisão, na medida em que se pode codificar qualquer função booleana. No entanto, uma vantagem das regras DNF é que estas tendem a gerar classificadores mais compactos do que a árvores de decisão. Métodos de regras aprendizagem geralmente tentam selecionar entre todas as possibilidades convergindo para a melhor regra de acordo com algum critério de minimalidade. Enquanto as árvores de decisão são construídas por uma estratégia *topdown*, a regras DNF muitas vezes são construídas de forma *bottom-up*. Inicialmente, todo o exemplo de treinamento d_j é visto como uma cláusula $\eta_1, ..., \eta_n \rightarrow \gamma_i$, onde $\eta_1, ..., \eta_n$ são os termos contidos em d_j e γ_i igual a *classe* ou a $\bar{classe}$ dependendo se d_j é um exemplo positivo ou negativo da *classe*. Esse conjunto de cláusulas já é considerado um classificador DNF para a *classe*, com pontuações mais altas em termos de *overfitting*.

O algoritmo de aprendizado aplica então um processo de generalização em que a regra é simplificada por meio de uma série de modificações que maximizam a sua compacidade, não afetando ao mesmo tempo a propriedade de cobertura do classificador. No final do processo, é aplicada a fase de poda, em essência semelhante à empregada nas árvores de decisão, onde a capacidade de correta classificação de todos os exemplos de treinamento pode ser reduzida a fim de garantir mais generalidade.

```
PART decision list
------------------

male_okapi_contagem > 51 AND
female_cosseno_soma <= 35.10398: MALE (1651.0/461.0)

female_cosseno_soma > 32.17109 AND
male_okapi_contagem <= 35: FEMALE (113.0/9.0)

female_cosseno_soma > 34.27261: FEMALE (888.0/235.0)

male_okapi_contagem <= 46: FEMALE (839.0/312.0)

female_cosseno_contagem > 43: FEMALE (592.0/284.0)

: MALE (77.0/27.0)

Number of Rules  :      6
```

Figura 3.4: Exemplo de classificador baseado em regras para a atribuição de perfis de autoria

3.2.7 Árvores

Classificadores de árvores, como árvores de decisão, por exemplo, J48.

Árvore de decisão é um classificador em que os nós internos são rotulados por termos, ramos que partem destes são rotulados por testes considerando o peso que o termo tem no documento de teste, e as folhas são rotuladas por categorias. Tal classificador categoriza um documento de teste d_j por recursividade, testando os pesos que os termos dos nós

internos rotulados têm no vetor $\overrightarrow{d}_j$, até que um nó folha seja atingido; o rótulo deste nó é então atribuído à d_j. Muitos desses classificadores usam representação binária de documentos, resultando em árvores binárias. Um exemplo de Árvore de Decisão é ilustrado na Figura 3.5. Existe uma série de pacotes padrão para aprendizagem de árvore de decisão. Entre os mais populares estão ID3 e C4.5.

```
J48 pruned tree
------------------

male_okapi_contagem <= 51
|   female_cosseno_contagem <= 48: MALE (352.0/167.0)
|   female_cosseno_contagem > 48: FEMALE (2004.0/649.0)
male_okapi_contagem > 51
|   female_cosseno_soma <= 35.10398: MALE (1651.0/461.0)
|   female_cosseno_soma > 35.10398: FEMALE (153.0/56.0)

Number of Leaves  :     4

Size of the tree :      7
```

Figura 3.5: Árvore de decisão equivalente à regra DNF

A lista completa de classificadores utilizados nos experimentos é mostrada no Apêndice. Embora muitos desses classificadores possuam parâmetros que podem ser configurados, neste experimento foram utilizados os parâmetros em sua configuração padrão.

4 EXPERIMENTOS

Este capítulo descreve os materiais e métodos utilizados em nossos experimentos (Seção 4.1), e os resultados obtidos (Seção 4.2).

Os experimentos realizados em nossa investigação destinaram-se a analisar a qualidade dos atributos descritos na Seção 3.1. Também se testou vários algoritmos de aprendizagem, comparando seus resultados com a acurácia dos melhores sistemas do PAN 2013.

4.1 Materiais e Métodos

O objetivo desta secção é descrever os materiais utilizados no experimento e os métodos por meio dos quais ele foi realizado.

4.1.1 Conjunto de dados

O conjunto de dados de treinamento é composto de 236k arquivos XML contendo *posts* de blogs, conforme ilustrado na Figura 4.1. O conjunto de dados de teste é menor e tem conteúdo semelhante. Estes conjuntos de dados foram coletados de repositórios abertos e públicos, como Netlog, pelos organizadores do PAN2013 e estão disponíveis em: http://pan.webis.de/. Os detalhes dos conjuntos de dados são apresentados na Tabela 4.1.

O gênero e a idade dos autores são conhecidos, permitindo a utilização de algoritmos de classificação supervisionados.

```
<author lang="es" gender="male" age_group="20s">
    <conversations count="1">
        <conversation id="8f913d9cfc78b645b7846ab0735ble7e"> un amigo es aquel q
            esta ahy tan cerca de uno ……un amigo es aquel q nunca te falla ……..es por
            eso q un amigo es un segudo yo <img class="smiley"
            src="http://www.pan.net/smilies/smile.gif" alt=":)"/>
        </conversation>
    </conversations>
</author>
```

Figura 4.1: Exemplo de arquivo de entrada XML.

Em uma etapa de pré-processamento, foram removidos os caracteres de escape, *tags* e as ocorrências repetidas de caracteres de espaço. Não foi feito *stemming* e remoção de *stopwords*. O objetivo era manter o máximo possível do estilo de escrita dos autores. Em

Tabela 4.1: Detalhes dos conjuntos de dados de treinamento e teste do PAN

Classe	Treinamento	Teste
Feminino	118297	12717
Masculino	118291	12718
10s	17200	1776
20s	85796	9213
30s	133592	14446

seguida, o *post* foi separado em *tokens* usando NLTK[1].

Uma vez que os dados de treinamento foram pré-processados, os 61 atributos descritos na Seção 3.1 são calculados. De modo a calcular os atributos baseados em RI, foi utilizado o Zettair[2], que é um motor de busca compacto e rápido desenvolvido pela RMIT University (Austrália). Ele executa uma série de tarefas de RI, como indexação e correlação. Zettair implementa vários métodos para ranquear documentos em resposta às consultas, calculando as métricas Cosseno e Okapi BM25.

Para os atributos de legibilidade, o código *readability*.c[3] foi utilizado.

Nenhuma normalização ou discretização foi realizada nos valores dos atributos. Todas as instâncias do conjunto de treinamento foram utilizadas e nenhuma tentativa de equilibrar as classes foi realizada.

4.1.2 Métricas de Avaliação

Ao comparar a classe do *post* em relação ao gênero e à idade de fato contra as instâncias rotuladas pelo classificador é possível calcular as métricas de avaliação. O PAN utilizou acurácia como medida de qualidade. Acurácia é a proporção de casos classificados corretamente. Um problema em potencial é que a acurácia não é uma métrica confiável, pois irá produzir resultados imprecisos se o conjunto de dados for desequilibrado (ou seja, quando o número de casos em diferentes classes varia muito), que é o caso do conjunto de dados para a idade. Assim, os resultados apresentados nesta dissertação, também incluem medida-F, que é a média harmônica entre precisão e revocação.

Estes são calculados como:

$$Precisão = \frac{VP}{VP+FP} \tag{4.1}$$

$$Revocação = \frac{VP}{VP+FN} \tag{4.2}$$

$$Medida-F = 2 \times \frac{Precisão \times Revocação}{Precisão + Revocação} \tag{4.3}$$

$$Acurácia = \frac{VP+VN}{VP+FP+FN+VN} \tag{4.4}$$

[1] http://www.nltk.org/

[2] http://www.seg.rmit.edu.au/zettair/

[3] http://tikalon.com/blog/2012/readability.c

Onde VP, VN, FP e FN representam verdadeiro positivo, verdadeiro negativo, falso positivo, e falso negativo, respectivamente. Por exemplo, caso fosse necessário calcular as métricas para a classe gênero feminino, VP seriam os casos do gênero feminino que foram corretamente classificados como tal, FP representam as instâncias do gênero masculino que foram classificados como feminino, e FN representam as instâncias do gênero feminino que foram classificados como masculino.

A figura 4.2 mostra um exemplo de matriz de confusão que contabiliza os acertos e os erros feitos por ums hipótese avaliada.

		Classe prevista	
		Positivo	**Negativo**
Classe real	**Positivo**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Negativo	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Figura 4.2: Exemplo de matriz de confusão

4.2 Resultados

Esta seção descreve a seleção de atributos (Subseção 4.2.1), os classificadores (Subseção 4.2.2), os atributos por grupo (Subseção 4.2.3), uma comparação da nossa abordagem com os resultados oficiais do PAN2013 (Subseção 4.2.4), o desempenho dos classificadores em termos de velocidade (Subseção 4.2.5) e uma breve discussão sobre as abordagens e os resultados oficiais do PAN2014 (Subseção 4.2.6).

4.2.1 Seleção de Atributos

Um método amplamente utilizado para seleção de atributos é classificar os atributos com base em seu ganho de informação no que diz respeito à classe. Ganho de informação mede a redução esperada na entropia (ou seja, a incerteza), considerando apenas o atributo e a classe.

Segundo ROOBAERT; KARAKOULAS; CHAWLA (2006), ganho de informação mede a quantidade de informação em bits sobre a predição de classe, se a única informação disponível é a presença de um atributo e a correspondente distribuição de classe. Concretamente, ele mede a redução esperada na entropia. Dado o conjunto de exemplos de treinamento S_X, X_i é o i-ésimo vetor de variáveis neste conjunto, $|S_{X_i=v}|/|S_X|$ é a fração de exemplos da i-ésima variável possuindo valor v:

$$Ganho\ de\ Informação(S_X, X_i) = H(S_X) - \sum_{v=values(X_i)}^{\frac{|S_{X_i=v}|}{|S_X|}} H(S_{X_i} = v) \tag{4.5}$$

com entropia:

$$H(S) = -p = (S)log_2 p + (S) - P - (S)log_2 p - (S) \tag{4.6}$$

Tabela 4.2: Melhores e piores atributos para idade e gênero de acordo com seu ganho de informação

Melhores Atributos			
Idade		Gênero	
0,089	20s_cosseno_contagem	0,038	female_cosseno_contagem
0,086	30s_cosseno_soma	0,038	male_cosseno_contagem
0,083	30s_cosseno_contagem	0,038	female_cosseno_soma
0,081	20s_okapi_contagem	0,033	male_okapi_contagem
0,077	30s_okapi_contagem	0,033	female_okapi_contagem
0,055	30s_okapi_soma	0,023	female_okapi_soma
0,055	20s_okapi_soma	0,019	male_okapi_soma
0,051	20s_cosseno_soma	0,018	20s_okapi_soma
0,051	10s_okapi_soma	0,018	female_cosine_média
0,050	<href_html>	0,017	female_okapi_média
Piores Atributos			
Idade		Gênero	
0,013	female_cosseno_contagem	0,001	10s_okapi_contagem
0,011	repetição_ponto_final	0,001	repetição_ponto_final
0,007	<img_html>	0,001	<img_html>
0,003	repetição_ponto_exclamação	0,000	repetição_ponto_interrogação
0,001	repetição_ponto_interrogação	0,000	repetição_letra_i
0,001	repetição_letra_a	0,000	repetição_letra_a
0,000	repetição_letra_u	0,000	repetição_ponto_e_vírgula
0,000	repetição_letra_i	0,000	repetição_ponto_exclamação
0,000	repetição_vírgula	0,000	repetição_vírgula
0,000	repetição_ponto_e_vírgula	0,000	repetição_letra_u

$p \pm (S)$ é a probabilidade de um exemplo de treinamento no conjunto S ser da classe positiva/negativa.

A Tabela 4.2 mostra os melhores e os piores atributos para idade e gênero, considerando o seu ganho de informação. Estes resultados mostram que os atributos baseados em RI são dominantes entre os mais discriminativos para idade e gênero. O número de links <href_html> também apresentou potencial discriminativo para idade. Vogais repetidas estão entre os atributos menos discriminativos, tanto para idade quanto para gênero. Como esperado, um atributo que foi projetado especificamente para a idade é um dos menos discriminativos para gênero (10s_okapi_contagem). No entanto, um atributo que foi projetado para ser útil para a idade (20s_okapi_soma) acabou sendo um bom discriminador para gênero.

Ganho de informação avalia atributos independentemente um do outro. No entanto, quando o objetivo é selecionar um bom conjunto de atributos, é desejável evitar os redundantes, mantendo aqueles que têm ao mesmo tempo uma alta correlação com a classe e uma intercorrelação baixa. Por exemplo, female_cosseno_contagem e male_cosseno_contagem foram os melhores atributos para previsão de gênero. No entanto, manter ambos pode representar um desperdício se eles têm uma intercorrelação alta (isto é, se eles tendem a concordar sempre sobre a classe a ser prevista). Com este objetivo, foi utilizado os avaliadores de subconjunto do Weka para selecionar bons conjuntos de atributos. Neste sentido usamos além do método Ganho de informação o método *Wrapper* do Weka (HALL et al., 2009).

Segundo KOHAVI; JOHN (1997), na abordagem *wrapper*, mostrada na Figura 4.3, a seleção de subconjunto de atributos é feita usando o algoritmo de indução como uma

caixa preta. O algoritmo de seleção de subconjuntos realiza uma pesquisa para um bom subconjunto, usando o próprio algoritmo de indução, como parte da função de avaliação. A acurácia dos classificadores induzidos é estimada usando técnicas de estimativa de acurácia. A abordagem *wrapper* realiza uma busca no espaço de parâmetros possíveis. A pesquisa requer um espaço de estado, um estado inicial, uma condição de término, e um motor de busca.

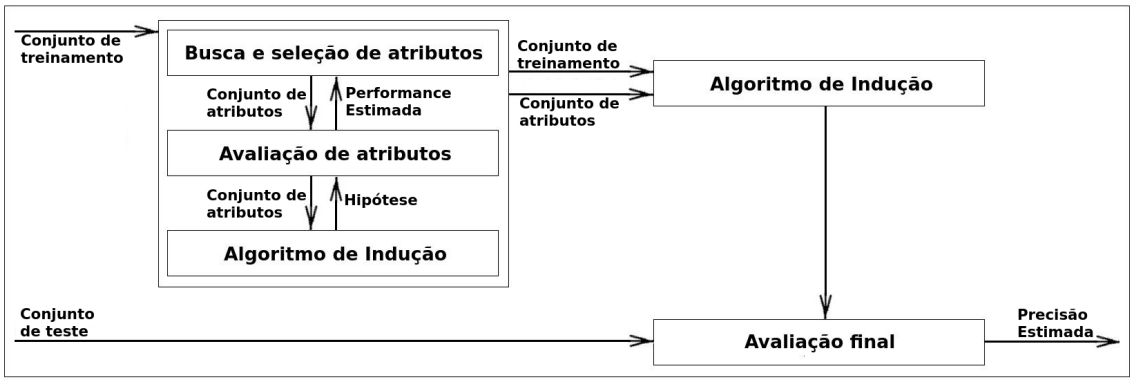


Figura 4.3: Abordagem *wrapper* para seleção de subconjunto de atributos (KOHAVI; JOHN, 1997)

A organização do espaço de busca é tal que cada estado representa um subconjunto de atributos. Para n atributos, existem n bits em cada estado, e cada bit indica se uma característica está presente (1) ou ausente (0). Operadores determinam a conectividade entre os estados, por padrão são utilizados os operadores que adicionam ou excluem um único atributo de um estado, que corresponde ao espaço de busca comumente usado em métodos passo a passo. A Figura 4.4 mostra como é o espaço de estados e operadores para um problema de quatro atributos. O tamanho do espaço de busca de atributos n é $O(2^n)$, de modo que não é prático pesquisar todo o espaço de forma exaustiva, a menos que n seja pequeno. O objetivo da busca é encontrar o estado com a maior avaliação, utilizando uma função heurística para guiá-la. Como a acurácia real do classificador induzido é desconhecida, a acurácia estimada é utilizada como função heurística e função de avaliação. Por padrão a função de avaliação que é utilizada é a de cinco vezes a validação cruzada (Figura 4.5) repetida várias vezes. O número de repetições é determinado em tempo real, olhando para o desvio padrão da acurácia estimada. Se o desvio padrão da acurácia é acima de 1% e cinco validações cruzadas não foram executadas, outra rodada de validação cruzada é executada. Embora esta seja apenas uma heurística, ela parece funcionar bem na prática e evita várias execuções de validação cruzada para grandes conjuntos de dados.

O subconjunto para gênero possui apenas seis atributos, a saber:

- female_cosseno_soma,
- male_cosseno_contagem,
- male_okapi_soma,
- male_okapi_contagem,
- <img_html>,

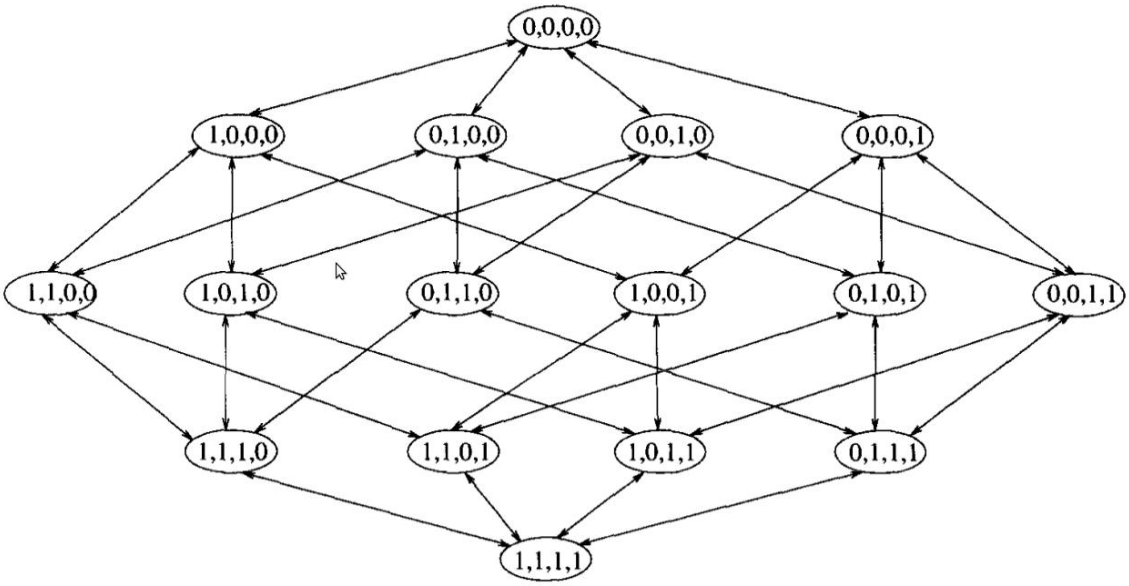


Figura 4.4: Busca no espaço de estados para seleção de subconjunto de atributos (KOHAVI; JOHN, 1997)

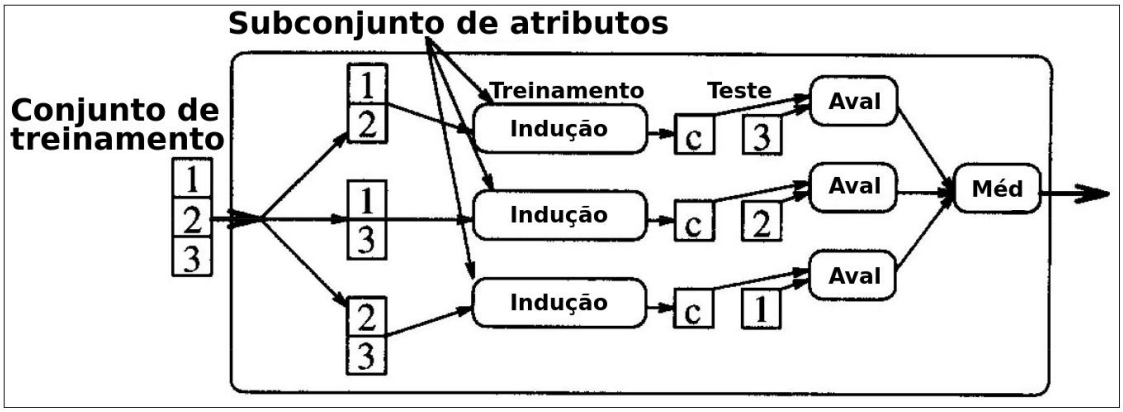


Figura 4.5: Método de validação cruzada para a acurácia estimada considerando 3 execuções (KOHAVI; JOHN, 1997)

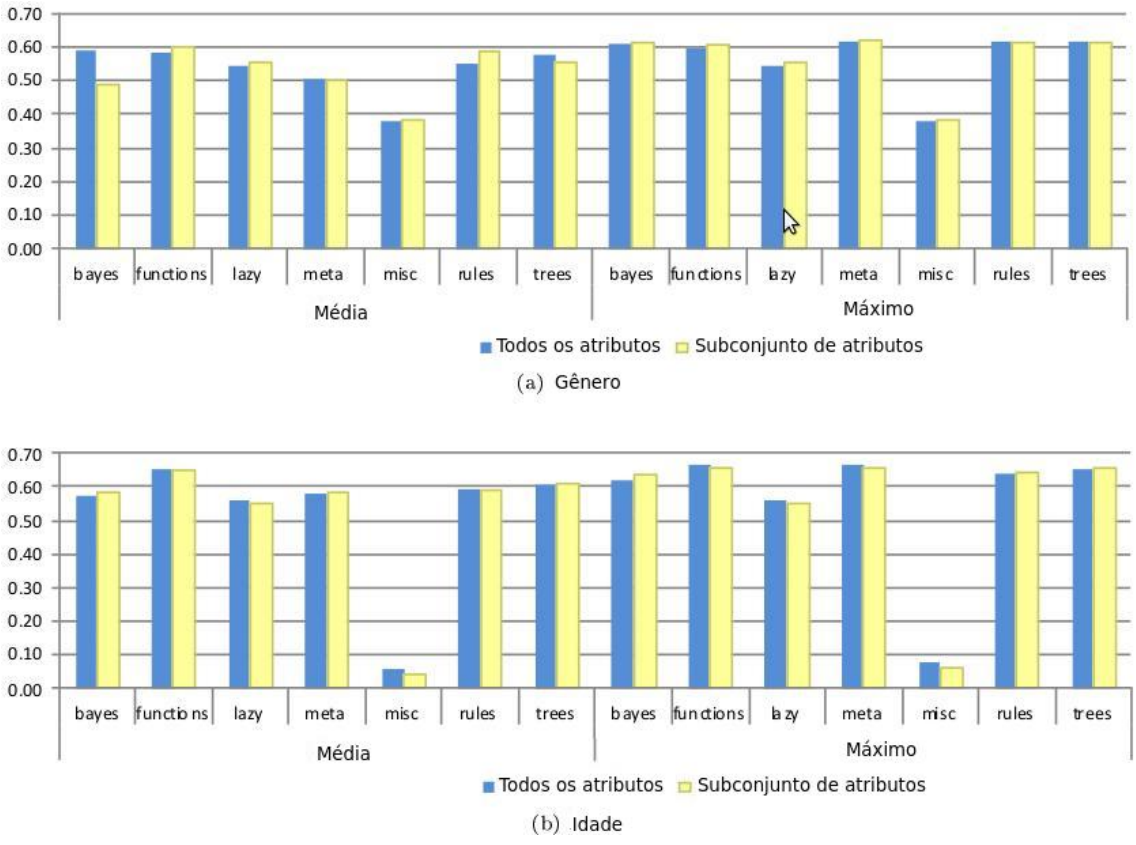
- limpeza.

O subconjunto para idade possui onze atributos, incluindo:

- 30s_okapi_média,
- 10s_cosseno_soma,
- 20s_cosseno_soma,
- 30s_cosseno_soma,
- 20s_cosseno_contagem,
- 30s_cosseno_contagem,
- 20s_okapi_soma,

- 20s_okapi_contagem,
- 30s_okapi_contagem,
- <href_html>,
- repetição_letra_e.

Foi realizada uma comparação das previsões dos classificadores executada em duas configurações (i), utilizando todos os 61 atributos e (ii) usando apenas o subconjunto de atributos. Os resultados são mostrados na Figura 4.6. Para a maioria dos algoritmos de aprendizagem, os resultados apresentados indicam diferenças pequenas entre os algoritmos testados, com uma ligeira vantagem em favor do teste executado com o subconjunto de atributos. A única exceção ocorreu com os classificadores Bayesianos para gênero, em que usar todos os atributos foi melhor. O teste T-pareado apresentou um valor-p de 0,27 para gênero e 0,77 para a idade, confirmando que não há nenhum ganho significativo no uso de todos os atributos. Neste caso, o uso do subconjunto é preferível, uma vez que o custo de computação e o treinamento do modelo é menor.



(a) Gênero

(b) Idade

Figura 4.6: Medidas-F máxima e média, considerando todos/subconjunto de atributos para gênero (a) e idade (b)

4.2.2 Classificadores

Dos 55 algoritmos de aprendizagem testados para gênero, a melhor pontuação em termos de medida-F foi apresentada pelo algoritmo *LogitBoost*. Além disso, observou-se que os resultados não têm grandes variações no que diz respeito ao tipo de algoritmos de

aprendizagem utilizados. Isto pode ser visto na Figura 4.6 e Tabela 4.3. Para a idade, o melhor algoritmo foi *Logistic* usando todos os atributos. Utilizando o subconjunto de atributos, verificou-se um empate entre *ClassificationViaRegression*, *RandomSubspace*, *MultilayerPerceptron* e *SimpleCart*. Isso mostra que os algoritmos de meta-aprendizagem, árvores de decisão, e os algoritmos de aprendizagem baseados em funções podem ser usados para atribuição de idade. No geral, os algoritmos de meta-aprendizagem apresentam melhores capacidades de previsão. Para gênero, as árvores e os algoritmos de aprendizagem Bayesiana também apresentaram bom desempenho.

Tabela 4.3: Performance em termos de velocidade dos dez melhores classificadores

Gênero

Classificador	Medida-F		Tempo (segundos)	
	Todos (61)	Subconjunto (6)	Todos (61)	Subconjunto (6)
meta.LogitBoost	**0,617**	0,615	18438	1896
meta.MultiBoostAB	0,616	**0,616**	9112	946
*trees.*J48	0,615	0,606	4718	280
rules.ConjunctiveRule	0,613	0,615	2156	175
trees.*DecisionStump*	0,610	0,615	**903**	**93**
meta.RotationForest	0,609	0,607	55885	3006
functions.MultilayerPerceptron	0,609	0,597	7936	195
rules.DecisionTable	0,607	0,600	1579	113
Bayes.BayesNet	0,606	0,587	984	95
trees.RandomForest	0,606	0,582	2881	1355

Idade

Classificador	Medida-F		Tempo (segundos)	
	Todos (61)	Subconjunto (11)	Todos (61)	Subconjunto (11)
functions.Logistic	**0,665**	0,653	141	**16**
meta.MultiClassClassifier	0,664	0,653	**139**	18
functions.SimpleLogistic	0,663	0,653	12128	988
meta.ClassificationViaRegression	0,661	**0,654**	7050	1314
functions.SMO	0,657	0,649	32148	1097
meta.RandomSubspace	0,656	**0,654**	586	79
meta.Dagging	0,655	0,649	3356	88
meta.RotationForest	0,655	0,652	58065	7169
functions.MultilayerPerceptron	0,654	**0,654**	10019	457
trees.SimpleCart	0,653	**0,654**	1835	1470

4.2.3 Atributos por Grupo

Além de analisar a utilidade dos atributos de forma individual, também é interessante avaliar a contribuição dos seis grupos de atributos descritos na Seção 3.1. Para este fim, os classificadores foram executados *n* vezes usando dois tipos de configurações: (i) remoção de um grupo de atributos de cada vez e (ii) manutenção de apenas um grupo de atributos de cada vez. Os resultados dessas execuções são mostradas na Tabela 4.4. Foram utilizados os classificadores com as mais altas medidas-F, considerando o conjunto completo de atributos, a saber, para gênero *meta.LogitBoost* e para idade *functions.Logistic*. Os piores resultados foram verificados quando os atributos baseados em RI foram removidos. Além disso, verificou-se que usando apenas este grupo de atributos, os classificadores tem melhores resultados do que quando se utilizaram todos os outros grupos. Surpreendentemente, descobrimos que os atributos de legibilidade (que têm sido amplamente utilizados

Tabela 4.4: Resultados da classificação removendo/mantendo cada grupo de atributos

Conjunto de Atributos		Gênero		Idade	
		Acurácia	**Medida-F**	**Acurácia**	**Medida-F**
Sem	Corretude	**0,621**	**0,614**	0,681	0,663
	RI	0,586	0,580	0,616	0,590
	Comprimento	**0,621**	0,613	0,675	0,657
	Legibilidade	**0,621**	0,613	**0,678**	0,660
	Sentimento	**0,621**	0,613	0,675	0,657
	Estilo	0,616	0,614	0,676	**0,658**
Somente	Corretude	0,552	0,535	0,601	0,550
	RI	**0,616**	**0,614**	**0,680**	**0,662**
	Comprimento	0,572	0,571	0,622	0,603
	Legibilidade	0,535	0,525	0,620	0,595
	Sentimento	0,554	0,533	0,614	0,592
	Estilo	0,573	0,559	0,628	0,608
Todos		**0,621**	0,613	0,682	**0,664**

em trabalhos relacionados) não só não ajudam, mas também reduziram a acurácia para a idade. Os outros grupos de atributos não apresentam impactos significativos na acurácia e medida-F.

4.2.4 Comparação com os Resultados Oficiais do PAN2013

A seguir, os resultados desta dissertação serão comparados com os resultados obtidos pelos participantes da tarefa de identificação de perfis de autoria que ocorreu no PAN 2013 (GOLLUB et al., 2013); (RANGEL et al., 2013). Esta comparação é mostrada na Figura 4.7, apresentando os resultados em termos de acurácia. Acurácia foi a métrica utilizada no PAN e uma vez que não temos acesso aos resultados em termos de precisão e revocação dos outros participantes, a medida-F não pode ser calculada. Os resultados apresentados por WEREN; MOREIRA; OLIVEIRA (2014) superam o desempenho do melhor grupo para idade e gênero considerando apenas o subconjunto de atributos descritos. Estes resultados superiores foram obtidos com um pequeno conjunto de atributos (seis por gênero e onze para a idade), que é muito menor do que o vencedor da competição (311 recursos para a idade e 476 para o sexo). Note-se que, apesar do gênero ter apenas duas classes, este apresentou uma dificuldade maior de previsão do que idade, que possui três classes. A maioria dos sistemas (incluindo o descrito nesta dissertação) tiveram escores mais elevados para a idade do que para gênero. Esta situação pode ser vista considerando os valores mais elevados para ganho de informação encontrados para a idade. Realizando agora uma comparação com os resultados do PAN2013, percebeu-se que usando mais atributos de RI e mantendo o mesmo algoritmo de classificação usado no PAN2013 (J48), a acurácia aumentaria 20%. Além disso, esses novos experimentos mostraram que o classificador (J48) usado na ocasião não está entre os melhores para atribuição de idade.

4.2.5 Desempenho em Termos de Velocidade de Processamento

O desempenho em termos de velocidade dos algoritmos de aprendizagem também é um aspecto muito importante. A tabela 4.3 mostra o tempo em que os dez melhores classificadores demoraram para gerar o modelo a partir dos dados de treinamento. Os melhores resultados para o tempo e a velocidade estão em negrito. Como esperado, o

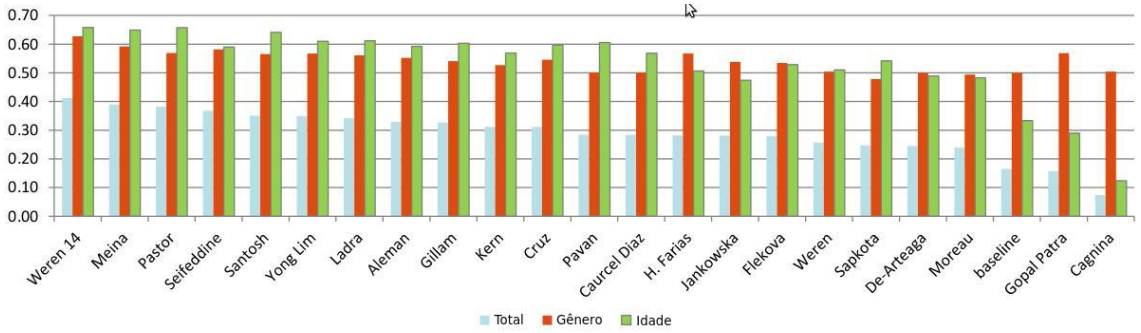


Figura 4.7: Comparação em termos de acurácia com os resultados oficiais do PAN

tempo necessário para construir o modelo utilizando todos os 61 atributos é muito maior do que quando se utiliza apenas um subconjunto. Isso é ainda mais perceptível para *MultilayerPerceptron*, para o qual a formação do modelo para gênero leva 40 vezes mais tempo quando se utilizam os seis grupos de atributos.

4.2.6 PAN2014

Esta subseção é baseada no *overview* realizado por RANGEL et al. (2014), avaliando as abordagens submetidas ao 11º laboratório de avaliação *Plagiarism, Authorship, and Social Software Misuse* (PAN)-2014, realizado como parte da *Conference and Labs of the Evaluation Forum* (CLEF). Assim como na edição do PAN2013, foi dada aos participantes a tarefa de identificação de perfis de autoria, mais especificamente identificar gênero e idade.

Dez equipes participaram da tarefa atribuição de perfis de autoria. Oito deles apresentaram artigos no formato de *notebook*, um (liau14) forneceu uma descrição de sua abordagem, e castillojuarez14 não relatou alterações em relação ao seu sistema apresentado no PAN2013.

A Tabela 4.5 mostra os atributos utilizados pelos participantes do PAN2014.

As acurácias mais elevadas foram atingidas por liau14 na identificação de gênero no twitter na língua inglesa (acurácia de 0,7338) e por MAHARJAN; SHRESTHA; SOLORIO (2014), na identificação de idade no twitter na língua espanhola (acurácia de 0,6111), bem como na identificação conjunta no twitter na língua espanhola (acurácia de 0,4333). É difícil estabelecer uma correlação entre as abordagens e os resultados, mas olhando para as três maiores acurácias por subcorpus e tarefa (sexo, idade e identificação conjunta), parece que atributos globais de conteúdo, tais como *bag-of-words* ou n-gramas alcançaram os melhores resultados. Da mesma forma, *bag-of-words* usados por liau14, n-gramas utilizados por MAHARJAN; SHRESTHA; SOLORIO (2014) e modelo vetorial usado por VILLENA-ROMÁN; GONZÁLEZ-CRISTÓBAL (2014) obtiveram os melhores resultados para quase todos os gêneros literários. Segundo (RANGEL et al., 2014) destaca-se também a nossa contribuição por meio da abordagem proposta nesta dissertação (WEREN; MOREIRA; OLIVEIRA, 2014), com atributos de RI usados para a identificação em blogs na língua inglesa; identificação conjunta nas mídias sociais na língua inglesa; identificação de idade no twitter; mídias sociais e reviews de hotéis na língua espanhola; identificação de gênero em blogs na língua espanhola e identificação conjunta nas mídias sociais na língua inglesa.

A combinação de atributos de conteúdo e estilo de MARQUARDT et al. (2014) rendeu bons resultados na identificação de gênero no twitter na língua espanhola e nas três iden-

Tabela 4.5: Atributos utilizados pelos participantes do PAN2014

	Autores								
Atributos baseados em estilo utilizados pelos participantes do PAN2014	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Frequência de sinais de Pontuação	X	X		X		X		X	
Tamanho da sentenças e palavras que aparecem uma e duas vezes além de desvios						X			
Número de caracteres e sentenças								X	
Calculo do número de posts por usuário, a frequência letras e palavras em maiúsculo					X				
Calculo da corretude, limpeza e diversidade do texto								X	
Calculo da ocorrência de tags HTML					X			X	
Índices de legibilidade	X	X			X	X		X	
Etiquetas morfossintáticas		X				X			
Emoticon				X	X				X
Atributos baseados em conteúdo utilizados pelos participantes do PAN2014	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
N-gramas									
Bag-of-words				X			X		X
Tópico das palavras						X			
Uso de MRC e LIWC para extração da frequência das palavras com diferentes conceitos psicolinguísticos					X				
Uso de dicionários para cada sub-corpus e classe	X								
Identificação de erros léxicos					X				
Identificação de palavras estrangeiras		X							
Identificação de expressões específicas					X				X
Atributos baseados em outras abordagens utilizados pelos participantes do PAN2014	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Atributos oriundos de sistemas de recuperação de informação								X	
Identificação de sentimento nas sentenças					X				
Representação de segunda ordem			X						

Legenda Autores:
1-(BAKER, 2014) **2**-(GILAD et al., 2014) **3**-(LÓPEZ-MONROY et al., 2014) **4**-(MAHARJAN; SHRESTHA; SOLORIO, 2014) **5**-(MARQUARDT et al., 2014) **6**-(MECHTI; JAOUA; BELGUITH, 2014) **7**-(VILLENA-ROMÁN; GONZÁLEZ-CRISTÓBAL, 2014) **8**-(WEREN; MOREIRA; OLIVEIRA, 2014) **9**-liau14

tificações (gênero, idade e em conjunto) em blogs na língua espanhola. A segunda melhor classificação obtida na identificação de gênero em mídias sociais na língua espanhola foi obtida com n-gramas, mas com rankings baixos nos outros subcorpora, demostrando que o uso de n-gramas não parece ser uma boa abordagem para atribuição de perfis de autoria em geral. O melhor desempenho geral foi obtido por LÓPEZ-MONROY et al. (2014), que empregaram representação de segunda ordem.

Considerando esta breve discussão sobre as abordagens submetidas ao PAN2014, é possível notar que a maioria das melhores abordagens utilizam atributos produzidos por operações dispendiosas. Neste sentido nossa abordagem traz uma clara vantagem pois utiliza atributos produzidos por operações simples com o uso de um sistema de RI.

Um diferença em relação ao PAN 2013 foi que além de identificar os perfis de autores de blogs, os participantes também foram convidados a construir abordagens para identificar os perfis de autores de Social Media, Twitter e Reviews de hotéis. A Tabela 4.6 mostra a distribuição de cada corpus de teste com relação à classe idade por idioma; com relação à classe gênero, todos os corpus eram equilibrados.

Tabela 4.6: Distribuição de cada corpus de teste com relação à classe idade por idioma (RANGEL et al., 2014)

	Social Media		**Blog**		**Twitter**		**Reviews**
	Inglês	**Espanhol**	**Inglês**	**Espanhol**	**Inglês**	**Espanhol**	**Inglês**
18-24	680	150	10	4	12	4	148
25-34	900	180	24	12	56	26	400
35-49	980	138	32	26	58	46	400
50-64	790	70	10	12	26	12	400
65+	26	28	2	2	2	2	294
TOTAL	3376	566	78	56	154	90	1642

O melhor desempenho geral foi obtido por LÓPEZ-MONROY et al. (2014), que obtiveram escore final de 0.2895, considerando a média das acurácias combinadas para cada subcorpus e idioma, conforme pode ser visto na Tabela 4.7.

Tabela 4.7: Resultados médios em termos de acurácia (RANGEL et al., 2014)

Time	**Média Geral**	**Média Inglês**	**Média Espanhol**	**Social Media**		**Blog**		**Twitter**		**Reviews**
				Inglês	**Espanhol**	**Inglês**	**Espanhol**	**Inglês**	**Espanhol**	**Inglês**
1	0,2895	0,2699	0,3156	0,1902	0,2809	0,3077	0,3214	0,3571	0,3444	0,2247
2	0,2802	0,2679	0,2967	0,1952	0,3357	0,2692	0,2321	0,3506	0,3222	0,2564
3	0,2760	0,2411	0,3226	0,2062	0,2845	0,2308	0,2500	0,3052	0,4333	0,2223
4	0,2349	0,2272	0,2452	0,1914	0,2792	0,2949	0,1786	0,2013	0,2778	0,2211
5	0,2315	0,2315	0,2316	0,1905	0,1961	0,3077	0,2321	0,2078	0,2667	0,2199
6	0,1998	0,1524	0,2631	0,1428	0,2102	0,1282	0,2679	0,1948	0,3111	0,1437
7	0,1677	0,1407	0,2037	0,1277	0,1678	0,1282	0,2321	0,1688	0,2111	0,1382
8	0,1404	0,1286	0,1563	0,0930	0,1820	0,0897	0,0536	0,1494	0,2333	0,1821
9	0,1067	0,0794	0,1430	0,1244	0,1060	0,0897	0,1786	0,0584	0,1444	0,0451
10	0,0946	0,1119	0,0716	0,1445	0,1254	0,1795	0,0893	-	-	0,1236
11	0,0834	0,1460	0,0000	0,1318	-	0,1282	-	0,1948	-	0,1291

Legenda Times:
1-lopezmonroy14 **2**-liau14 **3**-shrestha14 **4**-weren14 **5**-villenaroman14 **6**-marquardt14 **7**-baker14 **8**-baseline **9**-mechti14 **10**-castillojuarez14 **11**-ashok14

Os participantes que apresentaram software como em 2013, relataram suas experiências em trabalhos de *notebooks* que foram resumidos por RANGEL et al. (2014). Neste resumo foi realizada uma comparação pareada, considerando a acurácia de todos os sistemas, estabelecendo também os parâmetros de similaridade definidos na Tabela 4.8.

Tabela 4.8: Níveis de significância (RANGEL et al., 2014)

Símbolo	Nível de Significância
=	p > 0.05 não significante
*	0.05 > p > 0.01 significante
**	0.01 > p > 0.001 muito significante
***	p > 0.001 altamente significante

Tabela 4.9: Significância das diferenças de acurácia entre os pares de sistemas para identificação de gênero e idade em todo o corpus (RANGEL et al., 2014)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1		***	**	***	***	***	***	***	***	***	***
2			*	***	***	***	=	***	***	***	***
3				***	***	***	***	***	***	***	***
4					***	***	***	**	***	***	***
5						***	***	=	=	=	**
6							***	***	***	***	*
7								***	***	***	***
8									=	=	**
9										=	*
10											*
11											

Legenda Autores:
1-ashok **2**-baker **3**-castillojuarez **4**-liau **5**-lopezmonroy **6**-marquardt **7**-metchi **8**-shrestha **9**-villenaroman **10**-weren **11**-baseline

Tabela 4.10: Significância das diferenças de acurácia entre os pares de sistemas para identificação de idade em todo o corpus (RANGEL et al., 2014)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1		***	***	***	***	***	***	***	***	***	=
2			=	***	***	=	***	***	***	***	***
3				***	***	*	***	***	***	***	***
4					=	***	***	=	***	**	***
5						***	***	=	*	=	***
6							***	***	***	***	***
7								***	***	***	***
8									**	=	***
9										=	***
10											***
11											

Legenda Autores:
1-ashok **2**-baker **3**-castillojuarez **4**-liau **5**-lopezmonroy **6**-marquardt **7**-mechti **8**-shrestha **9**-villenaroman **10**-weren **11**-baseline

Com relação à identificação de idade, todos os sistemas são significativamente diferentes do *baseline*, exceto ashok14.

Existem alguns sistemas onde as diferenças não são estatisticamente significativas, tais como LÓPEZ-MONROY et al. (2014), liau14, VILLENA-ROMÁN; GONZÁLEZ-CRISTÓBAL (2014) e nossa contribuição através da abordagem proposta nesta dissertação (WEREN; MOREIRA; OLIVEIRA, 2014). Em blogs, a maioria dos sistemas são indistinguíveis, mas significativamente diferente do *baseline*. Nos outros subcorporas, a maioria dos sistemas também é diferente do *baseline*. Olhando para as acurácias, os resultados mostram que a maioria dos sistemas funcionam significativamente melhor do que o *baseline* na identificação de idade.

Tabela 4.11: Significância das diferenças de acurácia entre os pares de sistemas para identificação de gênero em todo o corpus (RANGEL et al., 2014)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1		***	**	***	***	***	***	***	***	***	***
2			*	***	***	***	=	***	***	***	***
3				***	***	***	***	***	***	***	***
4					***	***	***	**	***	**	***
5						***	***	=	=	=	*
6							***	***	***	***	*
7								***	***	***	***
8									=	=	*
9										=	*
10											*
11											

Legenda Autores:
1-ashok **2**-baker **3**-castillojuarez **4**-liau **5**-lopezmonroy **6**-marquardt **7**-mechti **8**-shrestha **9**-villenaroman **10**-weren **11**-baseline

Com relação à identificação de gênero, todos os sistemas são estatisticamente diferentes do *baseline*, mas LÓPEZ-MONROY et al. (2014), MARQUARDT et al. (2014), MAHARJAN; SHRESTHA; SOLORIO (2014), VILLENA-ROMÁN; GONZÁLEZ-CRISTÓBAL (2014) e nossa contribuição através da abordagem proposta nesta dissertação (WEREN; MOREIRA; OLIVEIRA, 2014) formam um grupo mais próximo. Em mídias sociais na língua inglesa, blogs nas línguas inglesa e espanhola e twitter na língua espanhola, a maior parte dos sistemas não apresentaram significativas diferenças estatísticas. Em mídias sociais para a língua inglesa todos os sistemas são diferentes do *baseline* e tiveram melhor desempenho na identificação de gênero. No twitter, a maioria deles apresentou melhor desempenho, porém para a mídias sociais na língua espanhola o contrário aconteceu e todos os sistemas tiveram um pior desempenho. O mesmo aconteceu em reviews de hotéis na língua inglesa, onde a maioria dos sistemas apresentou pior desempenho.

Observando as Tabelas 4.9, 4.10 e 4.11 e os parâmetros definidos na Tabela 4.8, é possível observar que a abordagem proposta nesta dissertação não possui diferença significativa em relação à abordagem que pontuou em primeiro lugar no PAN2014, confirmando que ela possui potencial para igualar e/ou superar o estado da arte.

5 CONCLUSÃO

Nesta dissertação, apresentamos uma avaliação empírica de uma série de atributos e algoritmos de aprendizagem para a tarefa de identificação de perfis de autoria. Mais especificamente, a tarefa abordada nesta dissertação refere-se a, dado um *post* de um blog, identificar o gênero e a faixa etária do seu autor. Esta é uma tarefa desafiadora, uma vez que os dados deste tipo de mídia apresentam muito ruído.

Realizamos experimentos extensos, considerando o ponto de referência criado para o PAN 2013. Seus resultados mostraram que os atributos baseados em RI podem levar a resultados que superam o estado da arte, além de resultados melhores que os vencedores do PAN 2013. Os atributos baseados em RI estão entre os mais discriminativos para idade e sexo.

Também comparamos nossa participação no PAN 2014 com as abordagens dos demais participantes. Tendo como conclusão a confirmação que nossa abordagem utilizando principalmente atributos derivados de RI possui potencial para igualar e/ou superar o estado da arte.

Na avaliação sobre os diferentes grupos de atributos, percebeu-se que qualquer grupo de atributos produz medida-F em torno de 0,55. Lembrando que esta é uma classificação binária (e o número de casos nas classes é equilibrado) e que um palpite aleatório produziria um resultado de 0,5. No entanto, passar de 0,6 é um desafio.

A inspeção visual dos dados mostra que os casos do gênero masculino e feminino se misturam e têm escores muito semelhantes para os atributos pesquisados. Esta situação incentiva o prosseguimento das pesquisas sobre atributos mais discriminativos para gênero. Descobriu-se também que atributos extraídos de um dicionário criado para análise de sentimentos não são úteis, e surpreendentemente o mesmo vale para os testes de legibilidade.

Um dos fatores que torna esta tarefa mais difícil é que os dados de treinamento são ruidosos. Uma vez que o conjunto de dados é composto de postagens de blogs, não há nenhuma exigência de que o conteúdo postado por seus autores seja gerado pelos mesmos. Neste sentido, cópia do conteúdo completo de outros autores é uma tendência comum. Por exemplo, encontramos 11 posts idênticos, oito deles foram marcados como masculino e três como feminino. Uma vez que a metodologia empregada baseou-se exclusivamente nos *posts* dos blogs, não há como distinguir diferentes classes em tais casos. Uma potencial solução envolve analisar outros atributos como o título e a URL do blog.

Em relação à escolha de classificadores, depois de testar os 55 algoritmos de classificação, não foi possível definir claramente o melhor para o domínio pesquisado, pois muitos deles renderam medidas-F semelhantes. Por exemplo, a classificação de gênero utilizando 30 dos 55 classificadores alcançou maior acurácia do que os vencedores do PAN2013. Esta constatação corrobora a sugestão de MANNING; RAGHAVAN; SCHüTZE (2008),

de que quando há uma grande quantidade de dados de treinamento, a escolha do classificador tem provavelmente pouco efeito sobre os resultados e a melhor escolha pode ser pouco clara. Nesses casos (isto é, quando a qualidade é semelhante), a escolha pode ser feita com base na escalabilidade. Neste contexto, esta escolha favorece o algoritmo *Logistc* para idade e os algoritmos *DecisionStump* e *BayesNet* para gênero.

Analisando nossos erros de classificação, percebemos que a maioria dos casos do grupo etário 10s foi erroneamente classificada como 20s e 30s. A revocação para esta classe foi muito baixa (cerca de 3%). Isso aconteceu porque a classe 10s teve poucos casos em comparação com as outras duas, o que introduziu um viés nos modelos de classificação.

Como mencionado na Seção 4.1, foram utilizados os parâmetros padrão para os algoritmos de aprendizagem. Os resultados obtidos aqui poderiam ser melhorados com o ajuste destes parâmetros, o que será realizado em trabalhos futuros.

Como trabalho futuro, pretendemos testar diferentes configurações para os classificadores que apresentaram os melhores desempenhos neste experimento, além de pesquisar métodos para seleção dos casos para o treinamento, visando à redução do ruído permitindo gerar modelos com melhor capacidade de previsão.

REFERÊNCIAS

ARGAMON, S. et al. Automatically profiling the author of an anonymous text. **Communications of the ACM**, [S.l.], v.52, n.2, p.119–123, Feb. 2009.

BAKER, C. I. Proof of Concept Framework for Author Profiling. In: NOTEBOOK FOR PAN AT CLEF. **Anais...** [S.l.: s.n.], 2014.

CARUANA, R.; NICULESCU-MIZIL, A. An Empirical Comparison of Supervised Learning Algorithms. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, New York, NY, USA. **Proceedings...** ACM, 2006. p.161–168. (ICML '06).

FAN, R.-E. et al. **LIBLINEAR - A Library for Large Linear Classification**. The Weka classifier works with version 1.33 of LIBLINEAR.

GILAD, G. et al. Ensemble Learning Approach for Author ProfilingâĂŤNotebook for PAN at CLEF 2014. In: NOTEBOOK FOR PAN AT CLEF. **Anais...** [S.l.: s.n.], 2014.

GOLLUB, T. et al. Recent Trends in Digital Text Forensics and Its Evaluation - Plagiarism Detection, Author Identification, and Author Profiling. In: CROSS-LANGUAGE EVALUATION FORUM. **Anais...** [S.l.: s.n.], 2013. p.282–302.

HALL, M. et al. The Weka data mining software: an update. **SIGKDD Explorations Newsletter**, [S.l.], v.11, n.1, p.10–18, Nov. 2009.

KINCAID, J. P. et al. **Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel.** [S.l.: s.n.], 1975.

KOHAVI, R.; JOHN, G. H. Wrappers for feature subset selection. **Artificial intelligence**, [S.l.], v.97, n.1, p.273–324, 1997.

KOPPEL, M.; ARGAMON, S.; SHIMONI, A. R. Automatically categorizing written texts by author gender. **Literary and Linguistic Computing**, [S.l.], v.17, n.4, p.401–412, 2003.

LÓPEZ-MONROY, A. P. et al. INAOE's participation at PAN'13: author profiling task. In: NOTEBOOK FOR PAN AT CLEF. **Anais...** [S.l.: s.n.], 2013.

LÓPEZ-MONROY, A. P. et al. Using Intra-Profile Information for Author Profiling. In: NOTEBOOK FOR PAN AT CLEF. **Anais...** [S.l.: s.n.], 2014.

MAHARJAN, S.; SHRESTHA, P.; SOLORIO, T. A Simple Approach to Author Profiling in MapReduce. In: NOTEBOOK FOR PAN AT CLEF. **Anais...** [S.l.: s.n.], 2014.

MANNING, C. D.; RAGHAVAN, P.; SCHüTZE, H. **Introduction to Information Retrieval**. New York, NY, USA: Cambridge University Press, 2008.

MARQUARDT, J. et al. Age and gender identification in social media. In: NOTEBOOK FOR PAN AT CLEF. **Anais...** [S.l.: s.n.], 2014.

MECHTI, S.; JAOUA, M.; BELGUITH, L. Machine learning for classifying authors of anonymous tweets, blogs and reviews. In: NOTEBOOK FOR PAN AT CLEF. **Anais...** [S.l.: s.n.], 2014.

MECHTI, S.; JAOUA, M.; BELGUITH, L. H. Author profiling using style-based features. In: NOTEBOOK FOR PAN AT CLEF 2013. **Anais...** [S.l.: s.n.], 2013.

MEINA, M. et al. Ensemble-based classification for author profiling using various features. In: NOTEBOOK FOR PAN AT CLEF. **Anais...** [S.l.: s.n.], 2013.

MOHAMMAD, S. M.; KIRITCHENKO, S.; ZHU, X. NRC-Canada: building the state-of-the-art in sentiment analysis of tweets. In: SEMANTIC EVALUATION EXERCISES (SEMEVAL-2013), Atlanta, Georgia, USA. **Proceedings...** [S.l.: s.n.], 2013.

MUKHERJEE, A.; LIU, B. Improving Gender Classification of Blog Authors. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, Stroudsburg, PA, USA. **Proceedings...** Association for Computational Linguistics, 2010. p.207–217. (EMNLP'10).

NGUYEN, D.; SMITH, N. A.; ROSÉ, C. P. Author Age Prediction from Text Using Linear Regression. In: ACL-HLT WORKSHOP ON LANGUAGE TECHNOLOGY FOR CULTURAL HERITAGE, SOCIAL SCIENCES, AND HUMANITIES, 5., Stroudsburg, PA, USA. **Proceedings...** Association for Computational Linguistics, 2011. p.115–123. (LaTeCH'11).

OTTERBACHER, J. Inferring Gender of Movie Reviewers: exploiting writing style, content and metadata. In: ACM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 19., New York, NY, USA. **Proceedings...** ACM, 2010. p.369–378. (CIKM'10).

PEERSMAN, C.; DAELEMANS, W.; VAN VAERENBERGH, L. Predicting Age and Gender in Online Social Networks. In: INTERNATIONAL WORKSHOP ON SEARCH AND MINING USER-GENERATED CONTENTS, 3., New York, NY, USA. **Proceedings...** ACM, 2011. p.37–44. (SMUC'11).

PENNEBAKER, J. W.; FRANCIS, M. E.; BOOTH, R. J. **Linguistic Inquiry and Word Count**. [S.l.: s.n.], 2001.

RAGHAVAN, S.; KOVASHKA, A.; MOONEY, R. Authorship Attribution Using Probabilistic Context-free Grammars. In: ACL 2010 CONFERENCE SHORT PAPERS, Stroudsburg, PA, USA. **Proceedings...** Association for Computational Linguistics, 2010. p.38–42. (ACLShort'10).

RANGEL, F. et al. Overview of the Author Profiling Task at PAN 2013. In: NOTEBOOK PAPERS OF CLEF LABS AND WORKSHOPS. **Anais...** [S.l.: s.n.], 2013. p.23–26.

RANGEL, F. et al. Overview of the 2nd Author Profiling Task at PAN 2014. In: NOTEBOOK FOR PAN AT CLEF. **Anais...** [S.l.: s.n.], 2014.

ROOBAERT, D.; KARAKOULAS, G.; CHAWLA, N. V. Information gain, correlation and support vector machines. In: **Feature Extraction**. [S.l.]: Springer, 2006. p.463–470.

SARAWGI, R.; GAJULAPALLI, K.; CHOI, Y. Gender Attribution: tracing stylometric evidence beyond topic and genre. In: FIFTEENTH CONFERENCE ON COMPUTATIONAL NATURAL LANGUAGE LEARNING, Stroudsburg, PA, USA. **Proceedings...** Association for Computational Linguistics, 2011. p.78–86. (CoNLL'11).

SEBASTIANI, F. Machine Learning in Automated Text Categorization. **ACM Computing Surveys**, [S.l.], v.34, n.1, p.1–47, Mar. 2002.

VILLENA-ROMÁN, J.; GONZÁLEZ-CRISTÓBAL, J.-C. DAEDALUS at PAN 2014: guessing tweet author's gender and age. In: NOTEBOOK FOR PAN AT CLEF. **Anais...** [S.l.: s.n.], 2014.

WEREN, E.; MOREIRA, V. P.; OLIVEIRA, J. Using Simple Content Features for the Author Profiling Task. In: NOTEBOOK FOR PAN AT CLEF. **Anais...** [S.l.: s.n.], 2013.

WEREN, E.; MOREIRA, V. P.; OLIVEIRA, J. de. Exploring Information Retrieval Features for Author Profiling. In: NOTEBOOK FOR PAN AT CLEF. **Anais...** [S.l.: s.n.], 2014.

WEREN, E. R. et al. Examining multiple features for author profiling. **Journal of Information and Data Management**, [S.l.], v.5, n.3, p.266, 2014.

WITTEN, I. H.; FRANK, E. **Data Mining**: practical machine learning tools and techniques. [S.l.]: Morgan Kaufmann, 2005.

ANEXO I

Tabela 5.1: Classificadores usados nos experimentos

lazy.IB1	functions.Logistic
lazy.IBk	functions.MultilayerPerceptron
meta.AdaBoostM1	functions.RBFNetwork
meta.AttributeSelectedClassifier	functions.SimpleLogistic
meta.Bagging	functions.SMO
meta.ClassificationViaClustering	misc.HyperPipes
meta.ClassificationViaRegression	misc.VFI
meta.CVParameterSelection	rules.ConjunctiveRule
meta.Dagging	rules.DecisionTable
meta.Grading	rules.DTNB
meta.LogitBoost	rules.JRip
meta.MetaCost	rules.OneR
meta.MultiBoostAB	rules.ZeroR
meta.MultiClassClassifier	trees.ADTree
meta.MultiScheme	trees.DecisionStump
meta.nestedDichotomies.ClassBalancedND	trees.J48
meta.nestedDichotomies.DataNearBalancedND	trees.LADTree
meta.nestedDichotomies.ND	trees.NBTree
meta.OrdinalClassClassifier	trees.RandomForest
meta.RacedIncrementalLogitBoost	trees.RandomTree
meta.RandomCommittee	trees.REPTree
meta.RandomSubSpace	trees.SimpleCart
meta.RotationForest	bayes.BayesianLogisticRegression
meta.Stacking	bayes.BayesNet
meta.StackingC	bayes.DMNBtext
meta.ThresholdSelector	bayes.NaiveBayes
meta.Vote	bayes.NaiveBayesMultinomialUpdateable
	bayes.NaiveBayesUpdateable