Emotional Fingerprint from Authors in Classical Literature

Matheus Araújo
Universidade Federal de Minas Gerais, Brazil
matheus.araujo@dcc.ufmg.br

Iuro Nascimento
Universidade Federal de Minas Gerais, Brazil
iuro@ufmg.br

Gustavo Caetano Rafael
Universidade Federal de Minas Gerais, Brazil
gustavorafael@dcc.ufmg.br

Raquel de Melo-Minardi
Universidade Federal de Minas Gerais, Brazil
raquelcm@dcc.ufmg.br

Fabrício Benevenuto
Universidade Federal de Minas Gerais, Brazil
fabricio@dcc.ufmg.br

ABSTRACT

The Internet deeply changed the way people share their knowledge. Almost all content that people produces is now available in digital formats, like e-books, apps, newspapers, and magazines. That content has commonly some metadata available that can be used to generate complex recommendation systems that track content similarity. Since there is some effort in the literature to explore this direction, almost all use classical recommendation approaches, like collaborative filter data and information present on websites that sells books. While most efforts in the literature use features derived from the text syntax to create a recommendation model, our approach aims to trace an emotional fingerprint of authors extracted from their texts. This approach, known as psychometry, consists of the study of behavioral characteristics like positivity, negativity, sadness, fear, religiosity, sexuality, which are able to disguise individuals. Using two sentiment analysis lexicons and a collection of 641 books from the English literature written by 56 authors, we show the effectiveness of these psychometric features in order to trace those authors emotional fingerprint.

Keywords

Authorship Identification; Sentiment analysis; Language features; Lexical semantics

1. INTRODUÇÃO

A Internet mudou profundamente as formas de publicação de conteúdo. Jornais e revistas são mais lidos hoje em sua forma online que na impressa e todos os tipos de textos agora estão disponíveis em formato digital. Em particular, a digitalização da língua escrita finalmente chegou a indústria editorial do livro através de e–books, que podem ser comprados e baixados a partir de vários tipos de livrarias online [8] [10]. E-books podem ser lidos em diferentes dispositivos eletrônicos como PDAs, Celulares e Tablets, além é claro de leitores de livros eletrônicos dedicados como o Kindle.

Essa mudança representa também um desafio e uma grande oportunidade para diversas disciplinas dentro da Ciência da Computação, em especial aquelas ligadas à descoberta de melhores métodos de processamento de linguagem natural e sistemas de recomendação. Como o conteúdo dos livros esta agora em sua forma digital, esses sistemas tem acesso a seus textos e metadados, que podem ser incorporados para compor mecanismos mais elaborados para recomendação de livros ou de novos autores. Apesar de existirem alguns esforços na literatura que exploram essa direção [22] [9] [25], a grande maioria é baseada em abordagens clássicas de recomendação, como filtragem colaborativa, e informações presentes em compras de livros realizadas online. Um exemplo é o site de compras Amazon que oferece recomendações do tipo: "Pessoas que se interessaram por esse livro se interessaram por esses outros". Nesse trabalho apresentamos uma abordagem inovadora capáz de identificar padrões textuais que proveem assinaturas características de autores em obras literárias. Nosso esforço consiste em explorar aspectos emocionais ao longo de todo o texto dos livros capazes de sumarizar unicamente a assinatura emocional de seu autor.

Nosso estudo foi baseado em técnicas de linguística forense, a ciência envolvida na identificação de autoria e utilizada por órgãos de segurança e inteligência a fim de identificar crimes de falsificações [20]. As análises realizadas neste tipo de estudo estão basicamente divididas em duas formas de análise, uma qualitativa e outra quantitativa. A forma qualitativa é limitada, pois trata da inferência do estilo e traços de escrita dos autores por peritos de forma subjetiva. Já a abordagem quantitativa é realizada a partir de uma avaliação objetiva das variações da escrita realizadas pelos autores através de ferramentas que auxiliam esta análise. Através da perspectiva qualitativa este trabalho propõe uma abordagem que pode ser incorporada nestas ferramentas de análise de autoria a fim de melhorar o seu desempenho.

A solução mais adotada para o problema de identificação de autoria na literatura envolve a identificação de classes de atributos estilométricos que se destacam na estrutura de um texto, esta solução é denominada estilometria. Estes atributos estruturais podem variar bastante, há trabalhos que utilizam a ocorrência de números, símbolos, formas de abreviações, formato do texto, variação na pontuação, ocorrência de verbos, advérbios e outras partes do discurso. Portanto como afirmado por [21], a estilometria busca atender aos diferentes pontos de vista de uma análise estilística forense,

WebMedia '16, November 08-11, 2016, Teresina, PI, Brazil

DOI: http://dx.doi.org/10.1145/2976796.2976868

incorporando em seu contexto, diferentes classes de atributos estilométricos a fim de identificar uma forma única da escrita de um autor.

Ao contrário da maioria dos trabalhos recentes na área que utilizam atributos derivados da sintaxe dos textos como ocorrência de verbos, adjetivos, outras partes do discurso, número de *stopwords*, pontuações dentre outros, a nossa proposta procura traçar uma assinatura emocional dos autores extraindo de seus textos somente atributos relacionados a fatores psicológicos como, por exemplo, positividade, negatividade, tristeza, medo, religiosidade, e sexualidade. Esta abordagem já é conhecida no campo da psicologia como psicometria, e consiste no estudo de características comportamentais capazes de distinguir sensorialmente indivíduos.

Através do uso da psicometria queremos responder questões de pesquisa como: "É possível atribuir a autoria de um conjunto de documentos apenas utilizando aspectos psicológicos presentes no texto? Existe uma assinatura emocional extraída a partir das palavras escolhidas pelo autor em sua obra? Qual é o desempenho desta técnica?". Para responder estas perguntas este trabalho propõe o uso de uma série de atributos psicométricos que foram selecionados a partir de 2 léxicos de sentimentos descritos na Tabela 1. Desta forma contribuímos na área ao enriquecer o leque de classes de atributos a serem utilizadas na análise estilométrica adicionando fatores psicológicos capazes de traçar uma assinatura emocional de cada autor da literatura clássica.

Para ilustrar as assinaturas emocionais propostas, apresentamos as Figuras 1 e 2. Os gráficos de radar ilustram 12 dos 62 atributos psicométricos utilizados neste trabalho, onde cada linha colorida representa um autor e os valores no gráfico representam as médias da ocorrência de determinado atributo ao longo de suas obras. Note que assim como impressões digitais, essa assinatura emocional visa identificar unicamente cada escritor. Os resultados demonstram que a representação emocional reflete a realidade literária de tais autores, como exemplo na Figura 1 podemos observar o Arcebispo Wake, autor que produziu muitos livros de cunho religioso [1], se distinguiu claramente em sua assinatura no quesito religião. Da mesma forma acontece com Marlowe, considerado por muitos um escritor com traços muito próximos a de Shakespeare [2], no gráfico sua assinatura é realmente muito próxima de Shakespeare apesar do fator sexualidade se destacar. Já na Figura 2, a assinatura de Charles Darwin apresenta poucas emoções destacadas, provavelmente devido ao caráter descritivo e científico de suas obras. Jules Verne, reconhecido por muitos como pai da ficção científica, possui em suas obras um equilíbrio emocional, sendo o atributo religião praticamente inexistente. Este resultado implica no questionamento sobre a religiosidade de Jules Verne, explorada em discussões como [5].

Existem muitas discussões e aplicações a serem exploradas para assinaturas emocionais. Portanto para novas reflexões desenvolvemos uma visualização online que permite comparar as assinaturas de cada autor extraídas deste trabalho dinamicamente.[3]

No entanto este trabalho procura dar um foco maior na validação desta técnica e entender o seu desempenho. As seções a seguir estão organizadas da seguinte forma, trabalhos relacionados para contextualizar este material, a metodologia adotada, os experimentos realizados para justificar a utilização de atributos psicométricos propostos e por fim a conclusão deste projeto.

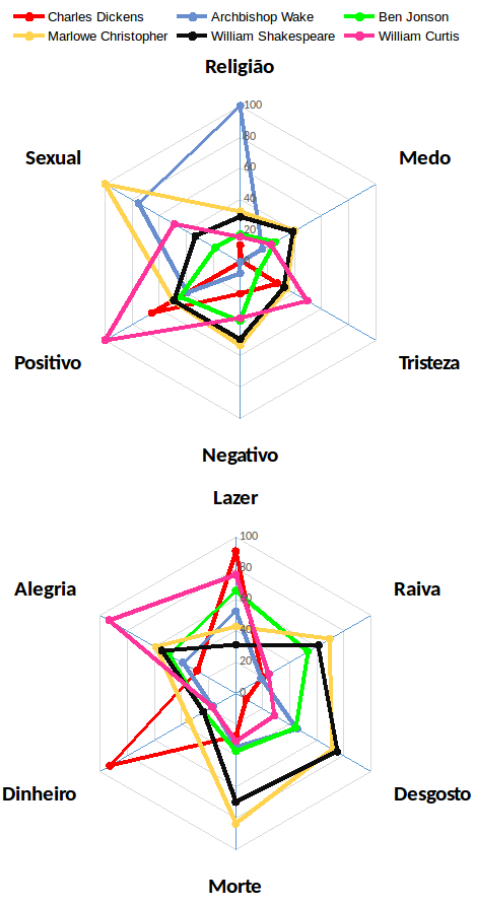

Figura 1: Exemplo da Assinatura Emocional proposta neste trabalho entre 6 autores e 12 diferentes atributos psicométricos. A assinatura emocional foi traçada a partir da média da ocorrência dos atributos psicométricos ao longo das diversas obras de cada autor.

2. TRABALHOS RELACIONADOS

A identificação de autoria em textos é um problema recorrente na literatura e apresentado de diferentes formas e contextos. Em [16], são apresentados diversos trabalhos realizados nesta área, sendo que três linhas de pesquisa possuem maior destaque: a identificação do autor, a descoberta de perfis dos autores e a detecção de plágio.

Dentre os muitos trabalhos realizados na área de atribuição de autoria de um texto, [16] é uma revisão recente, que apresenta publicações que focam em técnicas para identificação do perfil (idade, gênero e sexo) do autor, fornecendo uma ampla visão sobre a utilização de softwares como o dicionário LIWC (Linguistic Inquire Word Count) e de características de estilo e psicológicas da escrita.

Em [13] e [23] há trabalhos centrados em características estilométricas para atribuição do perfil dos autores de blogs, e-mails e Twitter. Nestes estudos foram extraídas as frequên-

[1]https://en.wikipedia.org/wiki/William_Wake
[2]https://en.wikipedia.org/wiki/Christopher_Marlowe
[3]Visualização Online: https://iurobpn.github.io/authorship/

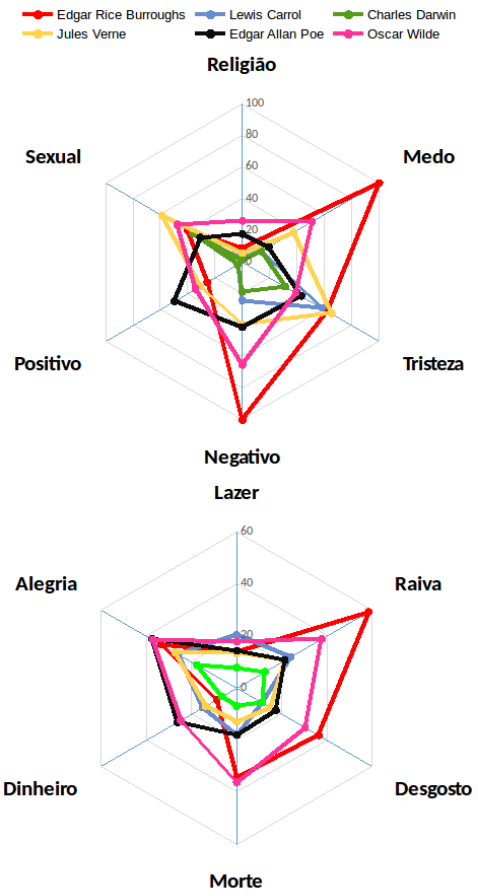

Figura 2: Assim como na Figura 1, um exemplo da Assinatura Emocional com outros autores famosos.

cias de palavras por categorias com o auxílio de dicionários como LIWC e são analisadas as características estruturais de texto como forma de reconhecer um autor. Nosso trabalho realiza algo semelhante, porém com o foco em apenas características psicométricas e em textos longos.

É relevante notar que a maioria dos trabalhos foca em características estilométricas para identificação e criação de um perfil dos autores e não na identificação de autoria propriamente dita. Em [3] e [23] algumas características psicológicas são incorporadas às análises do perfil dos autores mostrando um potencial nas características emocionais para identificação de assinatura de autores.

Mais recentemente, diversas novas técnicas foram incorporadas na identificação de autoria, como pode ser observado em [19]. Nesta revisão apresenta-se um apanhado dos trabalhos mais recentes da área, que em sua maioria foca na combinação de atributos para um algoritmo de aprendizado de máquina a fim de classificar autoria, assim como trabalhos que procuram selecionar quais atributos são os mais relevantes neste tipo de classificação. A maioria das técnicas utilizadas foca em características linguísticas e na estruturação sintática das frases.

Apesar dos significativos resultados na identificação dos autores de textos, trabalhos como [11] questionam a eficácia de métodos que utilizaram um pequeno número de autores em seus experimentos obtendo resultados próximos a 95% de acurácia. Segundo o autor, esses métodos possuem parâmetros superestimados (*overfit*), que não seriam capazes de manter tal desempenho caso o número de autores fosse aumentado. Além disso, há uma redução de desempenho à medida que o número de autores cresce: na identificação de 145 autores os melhores resultados são da ordem de 35% de acurácia. Em nosso trabalho procuramos ser generosos ao analisar 56 autores em um único experimento a fim de evitar este tipo de viés.

Outro ponto discutido é a dificuldade de identificar o autor com uma base de dados muito limitada como, por exemplo, tendo entre de cinco e dez mil palavras por autor. Uma solução para esse problema é aumentar a quantidade de textos de um autor. Em [24] foram selecionados autores da literatura clássica da língua inglesa para a identificação de autoria devido ao grande número de obras disponíveis. A partir de técnicas estilométricas, focando nas partes de discurso do texto e na estrutura sintática, os autores conseguiram resultados com cerca de 85% de acurácia na identificação de autoria.

Uma vez selecionadas as características a serem extraídas para análise do texto, a maioria dos autores utilizam técnicas de aprendizado de máquina para criar um modelo para classificação. Em [18] é apresentado um estudo completo das mais diversas técnicas utilizando caracterização de texto. Outros trabalhos como [4] comparam diferentes técnicas como PMC (Perceptron de Multiplas Camadas), árvores de decisão (*decision trees*) e o *Support Vector Machine* (SVM). A última, segundo o autor, se destaca por ser capaz de processar milhares de categorias eficientemente. Em [11] é utilizada a implementação do algoritmo SVM conhecida como Otimização sequencial mínima (SMO do inglês *Sequencial Minimal Optimization*) para lidar com o problema da limitação de dados.

Observamos que alguns trabalhos relacionados se esquivam da identificação da autoria e partem para assuntos como identificação de perfil de autores. Além disso, poucos buscam entender o impacto que os atributos psicométricos podem gerar para este tipo de tarefa. Portanto, nosso trabalho preenche estas lacunas inovando ao utilizar unicamente atributos psicométricos em diversos experimentos para identificação da autoria dos documentos.

3. METODOLOGIA

3.1 Conjunto de Dados

O Projeto Guntenberg[1] oferece mais de 38 mil livros clássicos gratuitamente para download, no entanto a utilização de todo este conjunto de dados para este trabalho se mostrou inviável por dois motivos: o primeiro seria a falta de estudos comparativos e o segundo é o alto índice de repetição de livros nesta base, o que poderia tornar este trabalho muito enviesado.

Para solucionar estes problemas, decidimos escolher como conjunto de dados ao longo deste trabalho os mesmos 56 autores e a mesma quantidade de livros utilizada por [24]. Todos os 56 autores estão listados na Tabela 2 e cada livro foi coletado manualmente do Projeto Gutenberg a fim de evitar que houvessem trabalhos repetidos na base. Foi também realizado um pré-processamento retirando os metadados e as licenças de copyright inseridas elo Projeto Gutenberg a fim de evitar ruídos na execução dos experimentos.

Devido a pouca cobertura de autores de diferentes épocas literárias pelo conjunto de dados descrito acima, em especial, para o experimento da Seção 4.4 utilizamos 5443 livros coletados automaticamente do projeto Gutenberg. Para evitar a não repetição de livros, utilizamos o identificador do livro presente em seus metadados para filtrar as entradas. Foi também realizada uma inspeção automática para identificar se o texto estava escrito na língua inglesa.

3.2 Métricas

Neste estudo utilizamos duas métricas de analise de desempenho o F-Measure e a Acurácia. A acurácia é definida pelo número de livros cujos autores foram corretamente atribuídos dividido pelo número de livros analisados. Já o F-Measure é média harmônica entre a precisão e o recall: $2 * \frac{(preciso*recall)}{(preciso+recall)}$. Seja tp, o número de livros positivos verdadeiros, tn positivos falsos, fp falsos positivos e fn falsos negativos. Temos que precisão é definida como: $\frac{tp}{tp+fp}$ e o recall é definido com: $\frac{tp}{tp+fn}$.

3.3 Atributos da Assinatura Emocional

Os atributos psicométricos que foram extraídos dos livros para identificar a assinatura emocional de autores são provenientes de dois léxicos de sentimentos, são eles o Emolex e o LIWC. Ambos estão entre vários léxicos de sentimentos na literatura. Em [17] são apresentados cerca de 19 métodos de análise de sentimentos e uma ampla comparação de desempenho entre eles. Em [2] é disponibilizado os métodos diretamente através de uma ferramenta online. No entanto, foram escolhidos o LIWC e o Emolex por possuírem vários atributos emocionais em seus léxicos não apenas a polaridade de sentimentos. Existem outras ferramentas com léxicos formados a partir de escalas psicométricas, como o PANAS-t [6], que deixamos para explorar em trabalhos futuros.

O Emolex [12], também conhecido como NRC Emotion Lexicon, contém mais de 14 mil palavras associadas a 10 emoções básicas mostradas na Tabela 1, estas emoções foram definidas através do trabalho de [15]. Já o LIWC (Linguistic Inquiry and Word Count) é uma ferramenta paga criada para realizar uma série de análises linguísticas em textos longos. A ferramenta calcula a ocorrência de 90 atributos linguísticos, porém apenas 52 estão relacionados a emoções, estes também podem ser visualizados na Tabela 1. Foi utilizada a versão 2015 do LIWC [14].

3.4 Identificação da Assinatura Emocional

Para a identificação da assinatura emocional de cada autor, avaliamos a ocorrência de cada atributo psicométrico ao longo de suas obras literárias. Essa extração é realizada através da contagem do percentual de palavras relacionadas a cada atributo presente na Tabela 1. Por exemplo, caso hajam 100 palavras relacionadas ao sentimento de tristeza (*sadness*) em um texto que possui no total 1.000 palavras, o percentual de ocorrência da tristeza é 10%. Devido à grande quantidade de palavras que não correspondem a nenhum atributo, os valores de incidência são baixos. Portanto, a fim de tornar mais fácil a identificação da assinatura emocional para cada autor foi realizada uma normalização Min-Max, desta forma é valorizada a diferença entre cada autor, sendo aquele possui o maior valor para um determinado atributo recebe 1 e o autor com a menor ocorrência de palavras ligadas a este atributo recebe 0.

3.5 Contribuição de cada Atributos nos Resultados

Na Figura 3, os atributos são apresentados de forma ordenada de acordo com o InfoGain calculado pela ferramenta WEKA [7] no conjunto de dados da Tabela 2. O atributo *focuspast* foi considerado o que possui o maior ganho de informação, sozinho é responsável por 4,5% do desempenho do F-Measure no experimento de identificação de autoria por validação cruzada. Pode-se observar também que a metade da quantidade de atributos deste trabalho, foi possível alcançar 0.6 de F-Measure para o máximo de 0,7 utilizando todas as features. Ao associar esta análise e a visualização no gráfico na Figura 3, podemos afirmar que apesar de cada atributo ter contribuído para a melhoria de desempenho, alguns atributos podem ser retirado a fim de economizar recursos ao analisar conjuntos de dados muito grandes.

3.6 Classificador SVM e parametrização

Para realizar a predição da autoria dos textos realizamos testes com classificadores como *Random Forest*, Perceptron de Múltiplas Camadas, dentre outros disponíveis no software WEKA. Este software disponibiliza gratuitamente a implementação de diversos métodos de aprendizado de máquina. Porém, como identificado por outros trabalhos relacionados, o SVM se destacou com o método de aprendizado com o melhor desempenho, portanto o SMO, uma variante do SVM na ferramenta WEKA foi utilizado ao longo deste trabalho.

O parâmetro C é uma constante que afeta complexidade do classificador SMO, e geralmente seu valor tem-se um *tradeoff* como desempenho do algoritmo. Portanto foi realizado uma busca por GridSeach pelo melhor valor do parâmetro C do SMO. Observamos que o aumento de C melhora a acurácia do algoritmo até $C \simeq 5,3$. Um aumento de C acima deste valor não afeta mais o desempenho.

4. EXPERIMENTOS E RESULTADOS

4.1 Identificação de Autoria

A tabela 2 mostra os resultados de 3 diferentes experimentos para a tarefa de identificação de autoria.

No experimento representado pela coluna **V. Cruz.** foi realizada uma validação cruzada de 10 dobras entre todos os conjuntos de livros dos autores. Nele 90% dos livros de cada autor foram utilizados para treinos e 10% dos livros foram separados para testes. Alguns autores possuem menos de 10 livros, nestes casos há um dificultador na identificação, pois sobram poucos livros para treino do classificador. Nota-se que os autores com mais de 20 livros têm F-Measure acima de $0,8$, enquanto a identificação de autores com poucos livros tem um desempenho pior. Apenas para os autores Marlowe, Poe e Bierce não foi possível identificar a autoria, sendo estes com menos de 10 livros para a análise. Em média foi alcançado 0,7 de F-Measure com um intervalo de confiança de 90% entre 0,63 e 0,77.

O segundo experimento representado pela coluna **LOO**, é um *Leave-One-Out*. Seja N o total de livros de um autor, $N-1$ livros são utilizados para treino e apenas 1 livro é deixado de fora. Desta forma N avaliações foram realizadas, um para cada livro, a fim de identificar como seria atribuída a autoria deste determinado livro caso tivéssemos uma base completamente rotulada e treinada pelo SVM. Podemos observar valores como 21/22 para Wodehouse, ou seja, 21 dos

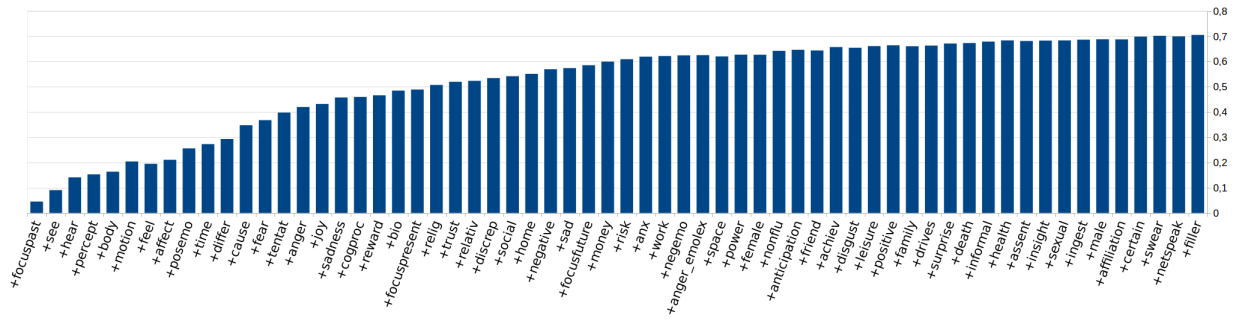

Figura 3: Desempenho acumulativo do F-Measure ao adicionar cada um dos atributos psicometricos na avaliação

22 livros escritos por Wodehouse foram corretamente classificados. Outros autores como Wakem, Alger, Fletcher, Alcott, Collins, também obtiveram bons resultados sendo suas assinaturas emocionais bem destacadas dos demais facilitando sua identificação. O experimento LOO possui resultados coerentes ao apresentado na coluna **V. Cruz.**, sendo que em média 64% dos livros foram classificados corretamente com um intervalo de confiança de 90% entre 0,57 e 0,72.

O terceiro experimento representado pela coluna **1 vs. T**, é chamado *todos contra um*. Seja A o número de autores, foram realizados um conjunto de A experimentos sendo que em cada experimento, remove-se a autoria de todos os livros com exceção dos livros de um determinado autor. Este experimento tem como objetivo analisar como a assinatura emocional dos autores funciona quando é desejado retirar os livros de um autor especifico de um conjunto de livros não rotulados. O classificador tem dificuldade neste experimento, pois aqueles livros que antes tinham autores definidos agora não os possuem, sendo assim as classes estão bastante misturadas, tornando mais difícil a identificação de um padrão para classificação. É interessante notar que neste experimento alguns autores que antes possuíam uma assinatura distinguível como Yonge, McCutcheon, Motley tiveram o F-measure reduzido drasticamente. No entanto, 43% dos livros foram classificados corretamente neste experimento com um intervalo de confiança de 90% entre 0,33 e 0,52.

4.2 Identificação Não Supervisionada

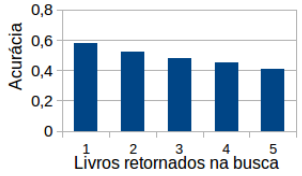

Figura 4: acurácia na busca por livros similares

A fim de, não apenas utilizar uma abordagem supervisionada como o o caso do classificador SVM, neste outro experimento foi avaliado como seria uma extração não supervisionada dos autores dado um determinado livro.

Este experimento foi inspirado no experimento P@5 realizado por [24], no qual o autor realiza através de um algoritmo de recuperação de informação uma consulta de um livro e recebe como resposta um ranking com os prováveis autores deste livro. Da mesma forma no nosso estudo, seja a assinatura emocional presente em um livro, uma consulta dessa assinatura é realizada por um algoritmo de recuperação de informação e são retornados como resultados os livros mais similares em forma de um ranking. A medida de desempenho utilizada foi a acurácia da busca de um livro de acordo com o autor do mesmo. Por exemplo, dado que foi pesquisado um livro de Shakespeare, se 3 dos 5 livros retornados são de Shakespeare, a acurácia é de 60%.

Para este experimento $N-1$ livros foram colocados em espaço F dimensional, sendo F a quantidade de *atributos* da Tabela 1. O livro deixado de fora é utilizado como *query*, no qual um algoritmo calcula os livros mais próximos utilizando a distância euclidiana.

Como resultado deste experimento temos o gráfico na Figura 4. Foram avaliados os 5 primeiros livros retornados na busca. Em média ao avaliar apenas o livro mais próximo, em 55% das vezes o resultado foi um livro do mesmo autor. O desempenho cai a medida que mais livros são considerados ao longo do ranking, com 5 livros apenas 40% dos livros retornados eram autor correto.

4.3 Desempenho Variando o Número de Classes

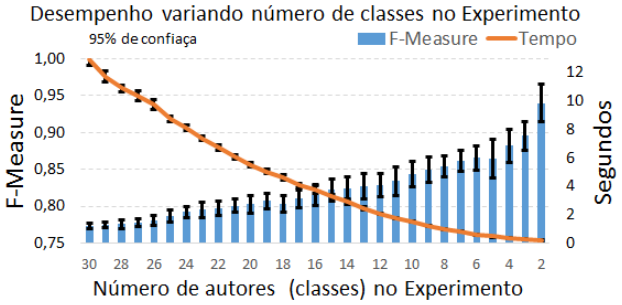

Figura 5: Apresentação do desempenho variando o número de classes no experimento para as métricas de tempo de execução e F-Measure. Observa-se que para 8 autores ou menos, o F-Measure é acima de 0,85 e tempo de execução abaixo de 5s.

Até agora, foram apresentados os resultados de experimentos avaliando 56 classes de autores com uma distribuição heterogênea de livros. No entanto gostaríamos de realizar um novo experimento a fim de identificar como a assinatura emocional se comporta em termos de tempo de execução e F-Measure variando o número de autores. Realizamos uma filtragem no conjunto de dados utilizando apenas autores que possuíam 10 ou mais livros, sendo que foram seleciona-

Tabela 1: Atributos Psicométricos Utilizados dos Léxicos Emolex e LIWC. Na coluna **Qnt** é apresentado o número de palavras no léxico associado ao respectivo atributo

Atributos do Emolex, Total: 14182 entradas		
Atributo	**Exemplos**	**Qnt**
Positive	abundance, shine, love	2312
Negative	abadon, death, nausea, hate	3324
Anger	doomsday, rage, savage	1247
Anticipation	inquiry, prognostic, prophecy	839
Disgust	prostitute, cholera, crap	1058
Fear	wilderness, threat, terrorism	1476
Joy	elegant, bless, amused	689
Sadness	punish, prison, fat	1191
Surprise	rarity, playful, mystery	534
Trust	radiance, proven, philosopher	1231
Atributos do LIWC 2015, Total: 6548 entradas		
Affective	processes affect happy, cried	1393
Positive	emotion posemo love, nice, sweet	620
Negative	emotion negemo hurt, ugly, nasty	744
Anxiety	anx worried, fearful	116
Anger	anger hate, kill, annoyed	230
Sadness	sad crying, grief, sad	136
Social	processes social mate, talk, they	756
Family	family daughter, dad, aunt	118
Friends	friend buddy, neighbor	95
Female	references female girl, her, mom	124
Male	references male boy, his, dad	116
Cognitive	processes cogproc cause, know, ought	797
Insight	insight think, know	259
Causation	cause because, effect	135
Discrepancy	discrep should, would	83
Tentative	tentat maybe, perhaps	178
Certainty	certain always, never	113
Differentiation	differ hasn't, but, else	81
Perceptual	processes percept look, heard, feeling	436
See	see view, saw, seen	126
Hear	hear listen, hearing	93
Feel	feel feels, touch	128
Biological	processes bio eat, blood, pain	748
Body	body cheek, hands, spit	215
Health	health clinic, flu, pill	294
Sexual	sexual horny, love, incest	131
Ingestion	ingest dish, eat, pizza	184
Drives	drives	1103
Affiliation	affiliation ally, friend, social	248
Achievement	achieve win, success, better	213
Power	power superior, bully	518
Reward	reward take, prize, benefit	120
Risk	risk danger, doubt	103
Past	focus focuspast ago, did, talked	341
Present	focus focuspresent today, is, now	424
Future	focus focusfuture may, will, soon	97
Relativity	relativ area, bend, exit	974
Motion	motion arrive, car, go	325
Space	space down, in, thin	360
Time	time end, until, season	310
Work	work job, majors, xerox	444
Leisure	leisure cook, chat, movie	296
Home	home kitchen, landlord	100
Money	money audit, cash, owe	226
Religion	relig altar, church	174
Death	death bury, coffin, kill	74
Informal	language informal	380
Swear	words swear fuck, damn, shit	131
Netspeak	netspeak btw, lol, thx	209
Assent	assent agree, OK, yes	36
Nonfluencies	nonflu er, hm, umm	19
Fillers	filler Imean, youknow	14

dos apenas 10 livros de cada autor para que não houvesse

Tabela 2: Lista de Autores e Desempenho na Identificação de Autoria em 3 experimentos de identificação de autoria utilizando SVM. Em **V. Cruz.** temos os resultados do F-Measure para o experimeto de classificação de autores utilizando validação cruzada. Em **LOO** temos os resultados para o experimento *Leave-One-Out*, os valores estão no formato *a/b*, onde a é são os livros classificados corretamente e *b*. Em **1 vs. T** temos a acurácia no experimento *um contra todos*.

Autores	**#Livros**	**V. Cruz.**	**LOO**	**1 vs. T**
Alcott	10	0,82	(9/10)	0,78
Alger	10	0,94	(15/16)	0,80
Austen	8	0,86	(6/7)	0,92
Baum	10	0,44	(4/10)	0,15
Bierce	8	0,00	(0/8)	0,43
Burroughs	9	0,94	(8/9)	0,89
Carroll	6	0,67	(4/6)	0,44
Churchill	22	0,83	(18/24)	0,47
Collins	23	0,93	(20/23)	0,91
Conrad	12	0,33	(4/11)	0,00
Curtis	7	0,67	(3/6)	0,67
Darwin	9	0,78	(7/9)	0,71
Defoe	9	0,24	(3/9)	0,20
Dickens	11	0,11	(2/9)	0,00
Fletcher	6	1,00	(6/6)	1,00
Galsworthy	10	0,80	(8/10)	0,89
Haggard	37	0,85	(30/33)	0,77
Hardy	7	0,53	(4/7)	0,00
Harte	9	0,67	(6/9)	0,50
Hawthorne	10	0,71	(6/10)	0,63
Henry	9	0,93	(13/14)	0,69
Holmes	9	0,48	(4/9)	0,00
Howells	10	0,44	(5/10)	0,00
James	19	0,75	(14/19)	0,09
Jonson	7	0,83	(5/6)	0,00
Kingsley	10	0,74	(6/10)	0,29
Kipling	8	0,50	(5/8)	0,00
Lang	10	0,31	(2/8)	0,00
Lever	9	0,82	(9/9)	0,31
London	21	0,80	(15/20)	0,65
Lytton	10	0,48	(4/10)	0,00
MacDonald	9	0,47	(5/9)	0,50
Marlowe	5	0,00	(0/5)	0,00
Maupassant	9	0,35	(4/9)	0,00
McCutcheon	10	0,89	(7/10)	0,00
Motley	10	0,77	(10/10)	0,22
Parker	10	0,43	(2/10)	0,27
Pepy	10	1,00	(10/10)	1,00
Poe	6	0,00	(0/6)	0,00
Rohmer	10	1,00	(10/10)	1,00
Schiller	10	0,70	(7/10)	0,63
Scott	10	0,90	(9/10)	0,90
Shakespeare	42	0,84	(37/42)	0,85
Shaw	10	0,76	(8/10)	0,71
Stevenson	10	0,27	(1/9)	0,00
Stockton	10	0,59	(5/10)	0,43
Tolstoy	15	0,57	(5/15)	0,32
Twain	14	0,50	(6/14)	0,00
Verne	10	0,74	(7/10)	0,46
Wake	9	1,00	(9/9)	0,94
Warner	10	0,46	(5/9)	0,27
Wells	10	0,70	(8/10)	0,43
Wilde	7	0,18	(0/7)	0,00
Wodehouse	23	0,93	(21/22)	0,93
Yonge	10	0,70	(7/10)	0,14
	Média	0,70	0,64	0,43
	D. Pad	0,27	0,29	0,36
	90 % IC	[0,63|0,77]	[0,57|0,72]	[0,33|0,52]

um viés, no total foram avaliados um conjunto de 300 obras de 30 autores.

Neste experimento de classificação, avaliamos o desempenho da validação cruzada de 10 dobras utilizando o classificador SVM. A primeira execução foi realizada para todos os 30 autores, em seguida novas execuções foram realizadas à medida que o número de autores no experimento era decrementado, sendo que o experimento termina quando há apenas 1 autor a ser classificado.

Na Figura 5, mostramos os resultados após 30 repetições aleatórias deste experimento com o respectivo intervalo de confiança de 95%. A linha no gráfico apresenta a média do tempo de execução e as colunas apresentam a média do F-Measure a medida que o número de autores e decrementado. Observa-se que avaliando 30 autores o F-Measure fica em torno de 0,77 e o tempo de execução 13s. A medida que o número de autores diminui observamos um rápido decréscimo no tempo de execução assim com um aumento significativo do F-Measure. Este comportamento mostra que a diminuição do número de autores diminui tanto a complexidade da tarefa de classificação quanto na quantidade de dados a serem avaliados pelo SVM. Com menos de 8 autores a serem avaliados temos um F-Measure médio acima de 0,85 e um tempo de execução abaixo de 5s.

Houve uma preocupação com este aspecto, pois em trabalhos como [11] é questionado o pequeno número de classes utilizadas nos experimentos deste tipo na literatura. Nota-se que quando há uma quantidade maior e homogênea de livros para cada autor, a assinatura emocional de cada autor se comporta muito bem ao identificar a autoria de livros.

4.4 Identificação de Épocas Literárias

Ao longo deste trabalho levantamos a hipótese de que não apenas é possível realizar a identificação de autoria através da assinatura emocional dos autores, mas também é possível traçar uma assinatura emocional das épocas literárias em que estes autores se encontram. Partimos do princípio que determinados estágios na história da literatura como Romantismo, Renascença, Iluminismo compartilham aspectos e perspectivas psicológicas similares dentro de suas obras.

De forma a obter a época literária a qual pertencem os autores e conseqüentemente os livros, foram utilizados os metadados do Projeto Gutenberg e um novo conjunto de livros e autores como explicado na Seção 3.1. O projeto tem um catálogo em arquivos de formato RDF[4], que é atualizado diariamente. Este catálogo oferece muitas informações úteis sobre autores e livros, como ano de nascimento e de morte do autor, gênero do livro e língua. As épocas foram selecionadas levando-se em consideração o ano de nascimento dos autores e uma janela de tempo de 20 anos. As classes de épocas foram definidas pelos seguintes intervalos de anos[5]:

- Medieval: $500 - 1500$;
- Renascença: $1500 - 1670$;
- Iluminismo: $1700 - 1800$;
- Romantismo: $1800 - 1820$.

Somente as épocas da literatura inglesa e livros em inglês foram levados em conta. A partir de 1820, as épocas começam a se entrelaçar e não seria possível classificar os livros com base no ano de nascimento do autor, assim somente o início do Romantismo foi considerado, de 1800 a 1820. Outro detalhe é que dado o período conturbado e de baixa produção literária na idade média, e a ausência de tecnologia como a prensa, há poucos livros desta época.

	Id. Méd.	**Ilumin.**	**Renac.**	**Romant.**
Id. Méd.	48	20	14	21
Ilumin.	1	1358	32	490
Renac.	4	81	505	20
Romant.	4	449	33	2363

Tabela 3: Matriz de Confusão por Movimento Literário. Nas linhas da matriz estão os livros rotulados de acordo com a linha e nas colunas da matriz estão os livros classificados pela assinatura emocional como a respectiva coluna. Na diagonal verificamos os livros classificados corretamente.

Na Tabela 3, o valor na linha i e coluna j corresponde a quantidade de livros do movimento literário da linha i atribuídos como a coluna j pela assinatura emocional. Os maiores erros foram classificações erradas entre o Romantismo e Iluminismo e vice-versa. A provável causa destes erros a proximidade temporal das épocas. Em geral 78,5% das épocas foram classificadas corretamente utilizando a assinatura emocional dos livros, este dado indica que podemos estender a utilização das assinaturas emocionais na classificação não somente de autores de livros, mas como épocas literárias.

5. LIMITAÇÕES DESTE ESTUDO

Apesar de nossos resultados sugerirem que a assinatura emocional é capaz de distinguir autores, sabe-se que determinados autores mudam seu estilo ao longo de suas obras, tornando este trabalho mais difícil. Um exemplo desta mudança pode ser observado na Figura 6, na qual são comparados alguns atributos emocionais dos autores Ben Jonson e Charles Dickens. Uma inspeção inicial permite perceber uma grande variação dos atributos entre as obras de Dickens, enquanto Jonson apresenta uma maior regularidade no conjunto dos livros analisados. Os resultados obtidos são uma indicação da existência de uma assinatura emocional mais característica no caso de Jonson. Esta assinatura é comprovada nos resultados da Tabela 2, na qual os valores de desempenho ao identificar os livros de Dickens são muito inferiores ao de Jonson. Em nossos experimentos, observamos que autores com alto desvio padrão na ocorrência dos atributos psicométricos tendem a serem mais difíceis de ser identificados.

6. CONCLUSÃO

Enquanto muitos trabalhos na literatura focam demasiadamente na utilização de atributos estilográficos para caracterização de autoria em obras literárias, este trabalho introduz uma nova perspectiva na forma de identificar autoria em textos longos. Observa-se que a partir de atributos psicométricos provenientes de 2 Léxicos de Sentimentos, obtêm-se resultados satisfatórios na tarefa de identificação de 56 autores em mais de 600 livros da literatura clássica. Ressaltamos o resultado de 0.7 em média de F-Measure para validação cruzada, valor elevado ao considerar a grande quantidade de autores para esta tarefa e a carência de livros para alguns indivíduos. Pretendemos explorar futuramente novas assinaturas emocionais focando agora na comparação de métodos

[4]Resource Description Framework

[5]Retirado de: www.online-literature.com/periods/timeline.php

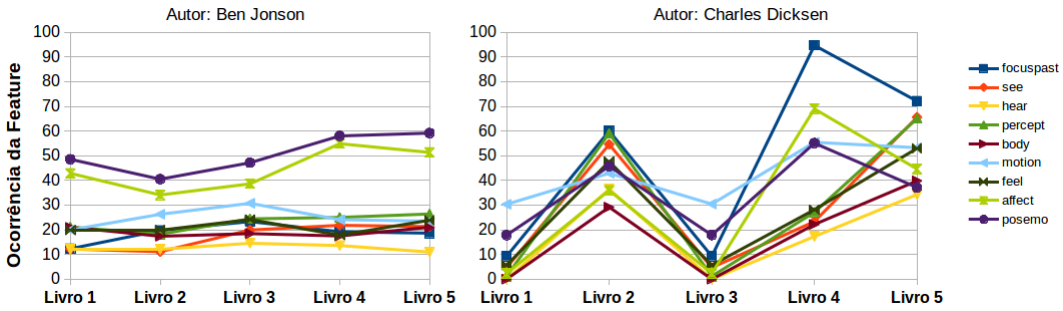

Figura 6: Evolução das Principais *Features* em Autores Distintos

estado da arte na identificação de autoria, dessa forma poderemos quantificar as melhorias que atributos psicométricos irão trazer para essas ferramentas.

Uma extensão do nosso estudo de autoria foi realizada a fim de identificar de épocas literárias por meio de da utilização da assinatura emocional. O trabalho foi capaz de identificar movimentos literários como Iluminismo, Renascentismo e Romantismo em 78,5% dos casos. Um trabalho futuro abordando a classificação de gêneros de livros como Romance, Terror, Suspense utilizando a assinatura emocional é merecido.

7. AGRADECIMENTOS

Financiado pelo projeto FAPEMIG-PRONEX-MASWeb, número do processo APQ-01400-14, e por bolsas de pesquisa individuais fornecidas pelo CNPq, CAPES, e Fapemig.

Referências

[1] Project gutenberg. http://www.gutenberg.org/.

[2] M. Araujo, J. P. Diniz, L. Bastos, E. Soares, M. Junior, M. Ferreira, F. Ribeiro, and F. Benevenuto. ifeel 2.0: A multilingual benchmarking system for sentence-level sentiment analysis. *ICWSM*, 2016.

[3] N. Cheng, R. Chandramouli, and K. Subbalakshmi. Author gender identification from text. *Digital Investigation*, 8(1):78–88, 2011.

[4] J. Diederich, J. Kindermann, E. Leopold, and G. Paass. Authorship attribution with support vector machines. *Applied intelligence*, 19(1-2):109–123, 2003.

[5] C. Frye. The religion and political view of jules verne, 2016.

[6] P. Gonçalves, F. Benevenuto, and M. Cha. Panas-t: A psychometric scale for measuring sentiments on twitter. *arXiv preprint arXiv:1308.1857*, 2013.

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[8] T. Hillesund. Will e-books change the world? *First Monday*, 6(10), 2001.

[9] Z. Huang, X. Li, and H. Chen. Link prediction approach to collaborative filtering. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 141–142. ACM, 2005.

[10] M. Levine-Clark. Electronic book usage: A survey at the university of denver. *portal: Libraries and the Academy*, 6(3):285–299, 2006.

[11] K. Luyckx and W. Daelemans. Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 513–520. Association for Computational Linguistics, 2008.

[12] S. Mohammad and P. D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29, 2013.

[13] R. Overdorf, T. Dutko, and R. Greenstadt. Blogs and twitter feeds: A stylometric environmental impact study.

[14] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn. The development and psychometric properties of liwc2015. *UT Faculty/Researcher Works*, 2015.

[15] R. Plutchik. *A general psychoevolutionary theory of emotion*, pages 3–33. Academic press, New York, 1980.

[16] T. R. Reddy, B. V. Vardhan, and P. V. Reddy. A survey on authorship profiling techniques. *International Journal of Applied Engineering Research*, 11(5):3092–3102, 2016.

[17] F. N. Ribeiro, M. Araújo, P. Gonçalves, M. A. Gonçalves, and F. Benevenuto. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):1–29, 2016.

[18] F. Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

[19] E. Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556, 2009.

[20] P. J. VARELA. *O uso de atributos estilométricos na identificação da autoria de textos*. PhD thesis, Pontifícia Universidade Católica do Paraná, 2010.

[21] P. J. Varela, E. J. Justino, L. E. Oliveira, and P. U. C. do Paraná. O uso de dicionário de atributos estilométricos na identificação de autoria de textos de língua portuguesa.

[22] P. C. Vaz, D. Martins de Matos, B. Martins, and P. Calado. Improving a hybrid literary book recommendation system through author ranking. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, pages 387–388. ACM, 2012.

[23] L. M. Werlen. Statistical learning methods for profiling analysis.

[24] Y. Zhao and J. Zobel. Searching with style: Authorship attribution in classic literature. In *Proceedings of the thirtieth Australasian conference on Computer science-Volume 62*, pages 59–68. Australian Computer Society, Inc., 2007.

[25] Z. Zhu and J.-y. Wang. Book recommendation service by improved association rule mining algorithm. In *Machine Learning and Cybernetics, 2007 International Conference on*, volume 7, pages 3864–3869. IEEE, 2007.