Brazilian Journal of Development

Um texto tão singular quanto a impressão digital: reconhecimento de autoria com um olhar para o Avasus

A text as singular as digital printing: author recognition with a look at Avasus

DOI:10.34117/bjdv5n12-352

Recebimento dos originais: 27/11/2019
Aceitação para publicação: 26/12/2019

Marcella Andrade da Rocha
Mestre em Engenharia Elétrica e Computação
Universidade Federal do Rio Grande do Norte – UFRN
E-mail: marcella.andrade@ufrn.edu.br

Roberto Douglas da Costa
Mestre em Ciência da Computação
Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte – IFRN
E-mail: douglas.costa@ifrn.edu.br

Ricardo Alexsandro de Medeiros Valentim
Doutor em Engenharia Elétrica e de Computação pela UFRN
Universidade Federal do Rio Grande do Norte – UFRN
E-mail: ricardo.valentim@ufrnet.br

Aline de Pinho Dias
Doutora em Educação pela Universidade Federal do Rio Grande do Norte - UFRN
Universidade Federal do Rio Grande do Norte – UFRN
E-mail: alinepinhodias@gmail.com

RESUMO

A Atribuição de autoria, a ciência de conferir a um autor determinado texto com base em suas características de escrita, é um problema com uma longa história. Neste artigo, tem-se como objetivo a apresentação de algumas técnicas de extração de características para o problema de atribuição e reconhecimento de autoria para fins de torná-lo uma ferramenta de uso na plataforma de Ensino a Distância do Ministério da Saúde, AVASUS. A literatura relevante apresenta as técnicas de análise de texto e características estilométricas de diversos textos que permitem que a autoria seja determinada. A extração das características servirá de base para uma pesquisa que tem como alvo o AVASUS, onde os estudantes de Ensino Superior na área de saúde fazem os cursos da plataforma, compartilham seus interesses e pensamentos em forma de mensagens nos fóruns e executam atividades que exigem redações sobre determinados temas na área da saúde. Essas produções escritas são o foco da aplicação da extração de características estilométricas para classificação e reconhecimento de autoria.

Palavras-chave: Extração de características; Ensino a distância; Ministério da Saúde; AVASUS; Reconhecimento de autoria.

Brazilian Journal of Development

ABSTRACT

Attribution of authorship, the science of giving an author certain text based on its writing characteristics, is a problem with a long history. This article aims to present some feature extraction techniques for the problem of attribution and recognition of authorship in order to make it a tool for use in the Distance Education platform of the Ministry of Health, AVASUS. The relevant literature presents text analysis techniques and stylometric characteristics of various texts that allow authorship to be determined. Feature extraction will serve as the basis for research targeting AVASUS, where higher education students take courses on the platform, share their interests and thoughts in the form of messages in the forums and perform activities that require writing. on certain health topics. These written productions are the focus of applying the extraction of stylometric features for classification and recognition of authorship.

Keywords: Feature extraction; Distance learning; Ministry of Health; AVASUS; Authorship Recognition.

1. INTRODUÇÃO

A atribuição de autoria exerce um papel importante em muitas aplicações, incluindo reconhecimento de autoria e investigação forense. As abordagens desse problema tentam identificar o autor de um documento por meio da análise do estilo de redação do indivíduo e/ou dos assuntos/tópicos sobre os quais ele costuma escrever. O problema tem sido extensivamente estudado e uma ampla gama de recursos tem sido explorada ((Hürlimann et al. 2015); (Stamatatos 2013); (Schwartz et al. 2013); (Seroussi et al. 2014)). Contudo, tem faltado a análise do comportamento das características em conjuntos de dados armazenados nos Ambientes Virtuais de Aprendizagem (AVA) ou usando uma série de classificadores. Consequentemente, fica difícil determinar quais tipos de características serão mais úteis para um determinado conjunto de dados no reconhecimento de autoria.

A atribuição de autoria é uma tarefa única que está intimamente relacionada à representação do estilo de escrita e à categorização do texto dos indivíduos. Em alguns casos, onde há uma distinção clara entre os documentos escritos por diferentes autores, as características relacionadas ao conteúdo, como aqueles usados na categorização de texto, podem ser eficazes. No entanto, é mais provável que as características baseadas em estilo sejam eficazes para conjuntos de dados que contêm um conjunto de conteúdo mais homogêneo.

Tradicionalmente, a tarefa sobre a atribuição de autoria de um texto é feita em um dos dois cenários: O primeiro é o da pesquisa literária e/ou histórica em que a atribuição é solicitada para um texto de origem desconhecida. Em complemento, onde, geralmente identificam autores em potencial, o trabalho é o reconhecimento de autoria, isto é, a seleção

de um autor em um conjunto de autores conhecidos; O segundo ambiente para atribuição de autoria em textos é o da linguística forense, onde precisa ser determinado se um suspeito escreveu ou não um texto específico, provavelmente incriminatório, onde a tarefa é a verificação de autoria que ocorre confirmando ou negando a autoria por um único autor conhecido (Halteren 2007).

Esse trabalho se concentra em parte nos dois cenários citados, a verificação e o reconhecimento de autoria na Educação a Distância (EaD), com o manuseio de um grande número de textos manipulados pelos estudantes no Ambiente Virtual de Aprendizagem do Sistema Único de Saúde - AVASUS.

A automatização do sistema EAD está iniciando e logo não será necessário um tutor para as atividades, apenas inteligência artificial. O sistema EAD mesmo sendo incorporado na internet não é completamente independente de intermediação humana de tutoria para ensino, correção e auxílio aos estudantes. O desenvolvimento de sistemas inteligentes está dominando o meio e automatizando diversos sistemas e isso envolve também o EAD, mas é um trabalho demasiadamente custoso, exige pesquisa em diversas áreas e será executado gradualmente. Com a automatização dos sistemas, as tarefas serão futuramente desempenhadas pelas máquinas através de inteligência artificial e implicará em redução de custos e qualidade por trazer resultados satisfatórios e segurança.

De acordo com o que possui hoje na plataforma AVASUS, melhorias são necessárias e devido a isso, a proposta do software do reconhecimento de autoria dos textos digitados pelos usuários nas atividades dos cursos irá enriquecer ainda mais o sistema e elevar a automatização do AVASUS, sendo assim um início para um sistema EAD totalmente mediado por tecnologia e inteligência artificial.

2. REFERENCIAL TEÓRICO

2.1 ATRIBUIÇÃO DE AUTORIA

A atribuição de autoria no campo científico foi consideravelmente desenvolvida, no decorrer da última década, aproveitando os avanços nas áreas da computação como aprendizado de máquina, recuperação de informação e processamento de linguagem natural. A diversidade de textos digitais disponíveis: mensagens de e-mail, blogs, fóruns on-line, códigos fonte, etc, aponta que a tecnologia existente, em virtude de uma ampla variedade de

aplicações, pode ser capaz de lidar com textos ruidosos de diversos autores candidatos. (Stamatatos 2011)

Em estudos recentes, como, Akimushkin et al. (2018), Albadarneh et al. (2015) e Al-Ayyoub et al. (2017) entre outros, grande parte dos algoritmos de Atribuição de Autoria é fundamentada em um modelo de representação simplificada usado no processamento de linguagem natural e recuperação de informações, conhecido como Bag of Words (BoW).

Vários estudos obtiveram resultados utilizando características léxicas, Altheneyan & Menai (2014) e Shojaee et al. (2013a), esse último utilizou Hápax legómenon (palavra registrada apenas uma vez) e Hápax dilegómenon (palavra registrada duas vezes) em um idioma.

Alguns trabalhos utilizando o Syntactic n-grams (sn-gram) (Sidorov et al. 2014), e o binary n-gram (Peng et al. 2016a), fazem variações ao método n-grama (sequência próxima de n itens de uma certa amostra de texto) para alcançar resultados melhores. Em particular, os n-gratengmas de caracteres são os mais populares devido a tolerância ao ruído e sua efetividade em documentos não-estruturados como e-mails, por exemplo. Embora os recursos n-gramas tenham se mostrado eficazes, a classificação baseada no mesmo é complexa (Brocardo et al. 2015). Contudo, n-gramas de caracteres nem sempre são melhores na precisão de classificação (Cerra et al. 2014). Em textos especialmente curtos a abordagem n-gramas de palavras se torna esparsa, pois a combinação das palavras não é encontrada, o que dificulta a classificação pelos algoritmos.

Destaca-se que, os trabalhos que foram abordados não englobam o reconhecimento de autoria como recurso para um Ambiente Virtual de Aprendizagem (AVA) e a utilização da língua portuguesa além da utilização de textos curtos (entre 70 e 300 caracteres) o que torna original a pesquisa exposta nesse artigo.

2.2 EDUCAÇÃO A DISTÂNCIA

Face as exigências do mercado, a EaD surge como um novo modelo educacional, com intuito de auxiliar na propagação do conhecimento de forma mais ágil, facilitando o acesso ao aprendizado.

O termo “Educação a Distância” pode ser conceituado por diversos espectros. Em (Brasil, 2005), a EaD, estabelecida no Art. 80 da Lei de Diretrizes e Bases da Educação Nacional de 20 de dezembro de 1996 e regulamentada pelo decreto lei nº 5.622 de 19 de dezembro de 2005, caracteriza-se como uma modalidade educacional que busca superar

limitações de espaço e tempo com a aplicação pedagógica de meios e tecnologias da informação e da comunicação e que, sem excluir atividades presenciais, organiza-se segundo metodologia, gestão e avaliação peculiares.

Já para Nunes (1994), a Educação a Distância pode ser considerada como um conjunto de ferramentas que possibilitam o atendimento de uma grande quantidade de alunos, independentes da localização geográfica e com alta qualidade, uma vez que não compromete o conteúdo e a forma de atendimento.

Em (Moore e Kearsley, 1996) a EaD é conceituada como uma metodologia de ensino e aprendizagem, facilitada por tecnologias, onde, os agentes envolvidos nesse processo, estão separados fisicamente e/ou temporalmente.

De acordo com Beloni (1999), a educação a distância está intensamente associada aos termos produção e qualificação, no qual as instituições de ensino, que trabalham com a Educação a Distância, investem cada vez mais em plataformas virtuais de aprendizagem para EaD em razão de sua flexibilidade, praticidade, custo baixo em relação ao ensino presencial e ambientes tecnológicos bastante adequados e facilitadores do processo ensino aprendizagem e capacitação docente.

Barros (2003) argumenta ainda que a educação a distância, além de ser um processo ensino e aprendizagem mediado por tecnologias, possui outras características a serem consideradas. Entre elas, destacam-se as diferenças de tempo e espaço, a necessidade de desenvolver hábitos para a autoaprendizagem.

Dessa forma, a EaD pode ser vista como uma metodologia de ensino, que permite a oferta de um ensino de qualidade, que supera o tempo e o espaço, e que está em constante atualização face as novas tecnologias que surgem, ao mesmo tempo que torna possível a incorporação dessas novas tecnologias ao processo de ensino e aprendizagem.

2.2.1 Plataforma Avasus

Hoje o cenário da EAD está diferente e mais pessoas podem ter acesso à educação de qualidade apenas com o uso da internet por meio dos AVA's, que é o mecanismo de distribuição do conteúdo dos cursos como também o local para interação entre aluno e professor (ABED 2016).

Sugestionado no modelo EAD foi criado o projeto AVASUS que consiste em uma plataforma de cursos EAD do Ministério da Saúde (MS), voltados para a capacitação de profissionais, professores, estudantes da área da saúde e também qualquer cidadão interessado

nos temas disponíveis. O conteúdo é feito por instituições de ensino superior e entidades ligadas a área a saúde que produzem módulos educacionais de qualificação e formação técnica conforme as necessidades da saúde pública no país (Vieira et al. 2017).

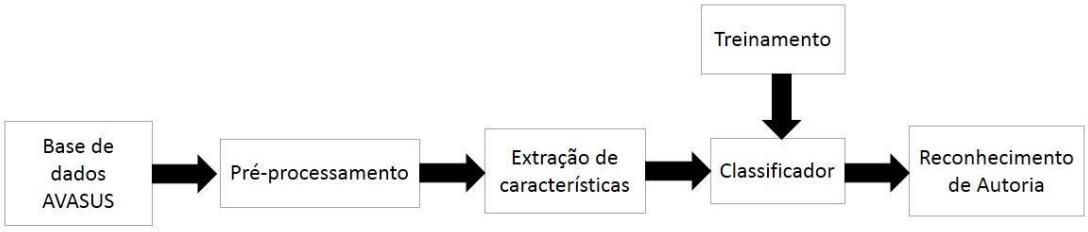
3. METODOLOGIA

O desenvolvimento do trabalho tem por finalidade a aplicabilidade dos conceitos para análise de textos e reconhecimento de autoria, mediante análise do padrão individual da escrita, como recurso para o AVASUS. Com isso, foi feito o pré-processamento dos dados. Extração de características e testes iniciais para classificação da autoria dos textos curtos escritos nos fóruns.

A etapa inicial constituiu em compreender as técnicas de análise de textos existentes na literatura e aplicá-las dentro do contexto de reconhecimento de autoria. Necessitou-se do banco de dados do AVASUS para aplicabilidade das técnicas. Foi escolhida a linguagem de programação Python por ser de alto nível e possuir uma gama de módulos para aplicação de Machine Learning, como o Scikit-learn, o módulo que integra uma gama variada de algoritmos para problemas supervisionados e não supervisionados de média escala (Pedregosa et al. 2011), que será utilizado para o software do sistema proposto.

O modelo do sistema proposto mostrado na figura 1 foi dividido em cinco partes: Base de dados do AVASUS; Pré-processamento; Extração de características; Treinamento e classificação; Reconhecimento de autoria.

Figura 1 - Etapas para o reconhecimento de autoria



3.1 PREPARAÇÃO DO DATAFRAME (BASE DE DADOS AVASUS)

A preparação e o carregamento dos *dataframe* do AVASUS são os primeiros passos para o modelo proposto. O objetivo é organizar o *dataframe* para a etapa de pré-processamento realizada em seguida.

A base de textos do AVASUS abrange alguns artigos de conclusão dos cursos e as respostas nos fóruns que os alunos publicam praticamente todos os dias no sistema. O número

de autores, comprimento do documento, quantidade vasta de entradas de textos com predisposição ao crescimento foram as características que influenciaram na seleção do banco de dados utilizado. A tabela 1 possui os dados do banco referente a quantidade de autores, total de textos e número de caracteres por documento em cada uma das configurações de teste.

Tabela 1 - Configurações de teste do banco de dados do AVASUS

Nº de testes	Teste 1	Teste 2	Teste 3	Teste 4
Gênero	Professores	Alunos	Professores	Alunos
Número de autores	11	11	11	14
Número de textos	3672	204	116	995
Número de caracteres (nc)	200<nc<800	40<nc<90	800<nc<3000	100<nc<400

3.2 PRÉ PROCESSAMENTO

O banco de dados do fórum do AVASUS possui 50939 registros de usuários únicos que publicam nos fóruns, para os testes iniciais foram utilizados 11 registros com cerca de 1000 publicações cada com mais de 70 e menos de 300 caracteres por texto, para balanceamento dos dados de treinamento e testes iniciais de reconhecimento de autoria, para futuramente expandir à sua totalidade e também pelo fato de que 50082 usuários fizeram de uma a dez publicações, o que não gera dados suficientes para treinamento e teste. Os dados são divididos em duas colunas: Identificação do usuário (**userid)** e Texto publicado no fórum, uma amostra do dataframe pode ser vista na tabela 2.

Tabela 2 - *Dataframe* do AVASUS

userid	**Texto**
44765	Boa noite.colegas. a icterícia fisiológica ou ...
5224	É uma etapa muito importante e necessária para...
35443	Acho muito importante o tema da ictericia neon...

24148	PARA A PREVENÇAO DA INFECÇAO POR HPV , SAÕ NEC...
109169	Boa noite na minha comunidade não utilizamos m...
2153	A promoçao e prevençao de saude da equipe na ...
43470	Evitar comportamentos sexuais de risco – Nomea...
5207	È um fato que no Brasil, aproximadamente 200 m...
124517	muito interessante e emocionante...
44765	Por exemplo e' de muita importância alimentaçã...
2710	Bom dia turma concordo com vocês sobre o acomp...
9120	O importante e evitar a doença, ese e o unico ...
54254	Boa noite Profa. gostaria de saber qual é a at...
9120	E é bem assim, ao redor de um 60 por cento dos...
45287	Boa noite pelas grandes vantagems que tem o al...
8994	Como se reconhece que uma criança está com ict...
24069	Resposta: Mito. é um problema cultural que vem...

Na etapa de pré-processamento para padronização do dataframe foram removidos os dados faltantes e nulos, tags HTML e espaços em branco. Logo em seguida realizou-se a redução dos dados para obtenção dos 11 registros, e remoção das StopWords, que são as palavras funcionais mais utilizadas na escrita como por exemplo os artigos, preposições, pronomes, etc., que não possuem muita relevância para o reconhecimento de autoria.

3.3 EXTRAÇÃO DE CARACTERÍSTICAS

Nesta etapa, já com os dados pré-processados, serão utilizadas técnicas de extração conhecidas que são relevantes para o reconhecimento de autoria como tokenização e vetorização.

3.3.1 Tokenização

Tokenização foi utilizada para converter as cadeias de texto normais em uma lista de tokens (palavras realmente relevantes) para classificação dos textos. A tabela 3 exibe um trecho dos dados convertidos em listas de *tokens*.

Tabela 3 - Dados convertidos em listas de Tokens

23009	[Bom, dia, educandos, educandas, É, prazer, te…
23583	[Agradeço, feedback, nome, equipe, AVASUS, Núc…
23867	[Percebo, claramente, participacao, pai, traz,…
24357	[Importante, sempre, convidar, pais, participa…
24478	[1, Tento, combater, refletindo, incentivando,…
25028	[Educandos, Berenice, concluiu, curso, Auricul…
25647	[Eva, obrigada, compartilhar, grupo, experiênc…

3.3.2 Vetorização

O conjunto de palavras utilizado podem ajudar a avaliar a (dis)similaridade entre os autores e expor a possibilidade do reconhecimento de autoria. Para isso, foi utilizado o "BoW" combinado com "Term Frequency-Inverse Document Frequency" (TF-IDF) (Ramos et al 2003), para definir quais palavras no corpus dos textos podem ser mais favoráveis de serem usadas pelo mesmo autor.

Assumindo o banco de dados normalizado e com os tokens (palavras realmente relevantes) extraídos. Foi aplicada duas etapas utilizando o modelo BoW:

1. Contar quantas vezes ocorre um *token* em cada mensagem (conhecida como frequência de termo);
2. Pesar as contagens, de modo que *tokens* frequentes recebem menor peso (frequência inversa do documento).

Primeiro passo, cada vetor terá tantas dimensões quanto houverem palavras únicas no corpo do texto. Em primeiro lugar será utilizado o *CountVectorizer* do *SciKit Learn*. Este modelo converterá uma coleção de documentos de texto em uma matriz de contagem de *token*.

Primeira etapa, cada vetor terá o tamanho da quantidade de palavras únicas no corpo do texto. Em primeiro lugar será utilizado o *CountVectorizer* do *SciKit Learn*. Este modelo converterá um conjunto de documentos de texto em uma matriz de contagem de *tokens*. Pode-

se imaginar isso como uma matriz bidimensional TF(t, $\underline{d}$). Onde a dimensão t é o vocabulário inteiro (1 *token* por linha) e a dimensão $\underline{d}$ são os documentos completos, neste caso uma coluna por mensagem de texto, criando assim uma matriz esparsa, como exemplo mostrado na tabela 4.

Tabela 4 - Matriz bidimensional

	Documento 1	**Documento 2**	**...**	**Documento N**
Token 1	0	1	...	0
Token 2	0	0	...	0
...	1	2	...	0
Token N	0	1	...	1

A matriz esparsa formada pelos dados após o BoW obteve o tamanho (linha, coluna) = (12018, 6914) e foi utilizada como entrada do classificador.

Na segunda etapa, para quantificar os *tokens* foi utilizado o *Inverse Document Frequency*, denotando o número total de textos como N definindo o IDF do texto de um token t como mostrado na equação 1, assim, o IDF de um token raro é alto, enquanto o IDF de um *token* frequente é baixo. Então o peso foi calculado como mostrado na equação 2 unindo as etapas 1 e 2.

$$log\ (\frac{N}{n_t}) \qquad (1)$$

$$W(t,\underline{d}) = tf(t,\underline{d}) \times log\ (\frac{N}{n_t}) \qquad (2)$$

Onde $W(t,\underline{d})$ é o peso do token t no documento d, $tf(t,\underline{d})$ é a frequência do token t no documento d, N é o número total de documentos, n_t é o número de documentos que contêm o termo t.

3.4 CLASSIFICAÇÃO E RESULTADOS

Nessa etapa de treinamento e classificação, utilizando dois algoritmos de *Machine Learning (NB e SVM),* por serem algoritmos clássicos e possuírem bons resultados de

classificação, *e* são feitos testes iniciais do reconhecimento de autoria. Inicialmente efetuou-se a divisão dos dados do conjunto em treinamento/teste, onde o modelo só "visualiza" os dados de treinamento durante a montagem e o ajuste de parâmetros. O tamanho do teste é de 20% do conjunto de dados inteiro (1360 mensagens do total de 6798), e o treinamento é o restante (5438 de 6798). Foi utilizado os recursos *pipeline* do *SciKit Learn* para armazenar uma linha de fluxo de trabalho. Isso permitirá configurar todas as transformações que serão feitas aos dados.

A tabela 5 descreve o desempenho dos dois classificadores, NB e SVM para a classificação da autoria dos 11 autores.

Tabela 5 - Desempenho dos classificadores

	Naive Bayes			*Support Vector Machine*		
	Precision	Recall	f1-score	Precision	Recall	f1-score
Micro AVG	0.83	0.83	0.83	0.95	0.95	0.95
Macro AVG	0.51	0.75	0.55	0.87	0.74	0.78
Weighted AVG	0.92	0.83	0.86	0.95	0.95	0.95

A tabela 5 mostra que, classificando os ID's dos textos do fórum, o SVM apresentou o melhor desempenho para todos os critérios entre os dois classificadores. O NB teve um desempenho inferior em comparação com o SVM. O classificador SVM teve uma precisão de 95% enquanto que o NB teve uma precisão de 83%. A boa eficiência adquirida pelos classificadores indicam uma maior (des)semelhança entre os autores no conjunto de dados, e, pode-se observar que os conjuntos de dados podem ter preferências de escrita claras entre os autores, o que faz com que o reconhecimento de autoria seja influenciado pela classificação do conjunto de palavras utilizado pelos autores.

4. CONCLUSÃO

Em conclusão, este estudo teve como objetivo investigar se seria viável o reconhecimento de autoria dos textos digitados nos fóruns do ambiente de aprendizagem virtual, AVASUS. Os resultados mostraram que técnicas de extração de características dos textos podem ser empregadas para identificar o autor de um texto e assim, futuramente, ser

utilizado para um número maior de ID's com a extração de uma quantidade superior de características. Entre os dois classificadores, o SVM exibiu o melhor desempenho, o que não descarta futuros testes com outros classificadores para uma quantidade maior de classes.

Com relação à limitação, mesmo que uma pessoa tenha uma impressão digital característica da escrita devido à sua maneira particular de aprender uma língua (H. Van Halteren et al 2005), as características que definem essa impressão digital são provavelmente complexas e não limitadas a uma única medida. Uma maneira possível de resolver essa limitação seria estender o método para empregar outras métricas.

O reconhecimento de autoria nas plataformas de Ensino a Distância é uma contribuição para automatização das plataformas ajudando a dificultar que terceiros resolvam as atividades no lugar dos usuários. Outra contribuição seria pelo fato de ser na língua portuguesa, a grande maioria dos trabalhos, são desenvolvidos na língua inglesa. Por fim, ficou perceptível que com a implementação futura do software inteligente será possível obter de forma automatizada o reconhecimento da autoria de um usuário do sistema.

REFERÊNCIAS

ABED (2016), Censo ead: Relatório analítico da aprendizagem a distância no brasil, Relatório técnico, Associação Brasileira de Educação a Distância, ABED, BR. URL: <http:// abed.org.br/ censoead2016/ Censo_EAD_2016_portugues.pdf>

AHMED, Al-Falahi, Ramdani Mohamed, Bellafkih Mostafa & Al-Sarem Mohammed (2015), Authorship attribution in arabic poetry, em 'Intelligent Systems: Theories and Applications (SITA), 2015 10th International Conference on', IEEE, pp. 1–6.

AKIMUSHKIN, Camilo, Diego R Amancio & Osvaldo N Oliveira Jr (2018), 'On the role of words in the network structure of texts: Application to authorship attribution 'Physica A: Statistical Mechanics and its Applications 495, 49–58.URL: <http://www.sciencedirect.com/science/article/pii/S0378437117312979>

AL-AYYOUB, Mahmoud, Yaser Jararweh, Abdullateef Rabab'ah & Monther Aldwairi (2017), 'Feature extraction and selection for arabic tweets authorship authentication', Journal of Ambient Intelligence and Humanized Computing 8 (3), 383–393.

ALBADARNEH, Jafar, Bashar Talafha, Mahmoud Al-Ayyoub, Belal Zaqaibeh, MohammadAl-Smadi, Yaser Jararweh & Elhadj Benkhelifa (2015), Using big data analytics

for authorship authentication of arabic tweets,em'Utility and Cloud Computing (UCC),2015 IEEE/ACM 8th International Conference on', IEEE, pp. 448–452.

ALTHENEYAN, Alaa Saleh & Mohamed El Bachir Menai (2014), 'Naïve bayes classifiersfor authorship attribution of arabic texts',Journal of King Saud University-Computerand Information Sciences26(4), 473–484.

BARROS, Daniela Melaré Vieira. (2003) "Educação a distância e o universo do trabalho", EDUSC, Bauru-SP.

BRASIL (2005). "Decreto 5.622 de19 de dezembro de 2005 regulamenta o Art. 80 da Lei no 9.394, de 20 de dezembro de 1996", 2005. Disponível em: http://portal.mec.gov.br/sesu/arquivos/pdf/portarias/dec5.622.pdf. Acesso em: 16 de Abr 2018.

BELLONI, Maria Luiza. (1999) "Educação a distância", Ed. Autores Associados, Campinas-SP.

BROCARDO, Marcelo Luiz, Issa Traore & Isaac Woungang (2015), 'Authorship verificationof e-mail and tweet messages applied for continuous authentication',Journal ofComputer and System Sciences81(8), 1429–1440

CERRA, Daniele, Mihai Datcu & Peter Reinartz (2014), 'Authorship analysis based on datacompression',Pattern Recognition Letters42, 79–84.URL:http://www.sciencedirect.com/science/article/pii/S0167865514000336

HALTEREN, H. Van, Baayen, F. Tweedie, M. Haverkort, A. Neijt, (2005) New machine learning methods demonstrate the existence of a human stylome, J. Quant. Linguist. 12 (1) 65–77.

MOORE, Michael G. & KEARSLEY, Greg (2011) "Distance Education: A Systems View of Online Learning (3rd ed.)", Wadsworth Cengage Learning, Belmont-US.

NUNES, Ivônio Barros. (1994) "Noções de educação a distância" Revista de Educação a Distância, Brasília, DF, p. 7–25.

PEDREGOSA, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg et al. (2011), 'Scikit-learn: Machine learning in python', Journal of machine learning research 12 (Oct), 2825–2830.

PENG, J., R. K. Choo & H. Ashman (2016a), Astroturfing detection in socialmedia: Using binary n-gram analysis for authorship attribution,em'2016 IEEE Trustcom/BigDataSE/ISPA', pp. 121–128.

RAMOS, Juan et al. (2003), Using tf-idf to determine word relevance in document queries, em 'Proceedings of the first instructional conference on machine learning', Vol. 242, Piscataway, NJ, pp. 133–142.

SHOJAEE, S., M. A. A. Murad, A. B. Azman, N. M. Sharef & S. Nadali (2013a), Detecting deceptive reviews using lexical and syntactic features,em'2013 13th International Conference on Intellient Systems Design and Applications', pp. 53–58.

SIDOROV, grigori, francisco velasquez, efstathios stamatatos, alexander gelbukh &liliana chanona-hernández (2014), 'syntactic n-grams as machine learning features fornatural language processing',expert systems with applications41(3), 853–860

STAMATATOS, Efstathios (2011), 'Plagiarism detection using stopword n-grams', Journal of the American Society for Information Science and Technology 62, 2512 – 2527.

VIEIRA, Geir Veras, Natanael de Freitas Neto, Karla Mônica Dantas Coutinho, Lidyane Alves da Cunha Laranjeiras, Ricardo Alexsandro de Medeiros Valentim & Karilany Dantas Coutinho (2017), 'Uma metodologia para otimizar o sistema de melhoria continuada do avasus com foco nas experiências do usuário', Revista Brasileira de Inovação Tecnológica em Saúde ISSN: 2236-1103 6 (3). URL: <https://periodicos.ufrn.br/reb/ article/view/11129>