UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
UNIDADE ACADÊMICA DE SISTEMAS E COMPUTAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

GABRIEL SOUTO MARACAJÁ

CLASSIFICAÇÃO DE GÊNERO AUTORAL BASEADO NA SUBJETIVIDADE DA LINGUAGEM

CAMPINA GRANDE - PB

2019

GABRIEL SOUTO MARACAJÁ

CLASSIFICAÇÃO DE GÊNERO AUTORAL BASEADO NA SUBJETIVIDADE DA LINGUAGEM

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Professor Dr. Claudio Elízio Calazans Campelo.

CAMPINA GRANDE - PB

2019

M298c Maracajá, Gabriel Souto.
Classificação de gênero autoral baseado na subjetividade da linguagem. / Gabriel Souto Maracajá. – 2019.

10 f.

Orientador: Prof. Dr. Claudio Elízio Calazans Campelo.
Trabalho de Conclusão de Curso - Artigo (Curso de Bacharelado em Ciência da Computação) - Universidade Federal de Campina Grande; Centro de Engenharia Elétrica e Informática.

1. Classificação de gênero autoral. 2. Linguagem subjetiva. 3. Léxicos. 4. Escrita subjetiva. 5. Gênero autoral. I. Campelo, Claudio Elízio Calazans. II. Título.

CDU:004(045)

Elaboração da Ficha Catalográfica:

Johnny Rodrigues Barbosa
Bibliotecário-Documentalista
CRB-15/626

GABRIEL SOUTO MARACAJÁ

CLASSIFICAÇÃO DE GÊNERO AUTORAL BASEADO NA SUBJETIVIDADE DA LINGUAGEM

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Professor Dr. Claudio Elízio Calazans Campelo
Orientador – UASC/CEEI/UFCG

Professor Dr. Herman Martins Gomes
Examinador – UASC/CEEI/UFCG

Professor Dr. Tiago Lima Massoni
Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 25 de novembro 2019.

CAMPINA GRANDE - PB

Classificação de Gênero Autoral Baseado na Subjetividade da Linguagem

Gabriel Souto Maracajá (Aluno), Cláudio Campelo (Orientador)
Departamento de Sistemas e Computação
Universidade Federal de Campina Grande
Campina Grande, Paraíba - Brasil

RESUMO

Identificar se um texto foi escrito por um homem ou por uma mulher é uma tarefa desafiadora. Porém, de grande relevância em diversas aplicações, desde a venda direcionada de produtos até campanhas de serviços de utilidade pública. Diferente dos diversos trabalhos que utilizam *Part of Speech*, com a finalidade de classificar gênero, neste trabalho, adotamos o grau de subjetividade dos textos como base para a classificação, partindo da suposição de que as mulheres escrevem de maneira mais subjetiva do que os homens. Para mensurar o grau de subjetividade, aplicamos técnicas que calculam a distância semântica dos textos à léxicos que representam várias dimensões de subjetividade. Nossos resultados mostram que homens e mulheres escrevem de forma correspondente, dificultando a ação dos algoritmos de classificação. Porém, acreditamos que nossos achados contribuem significativamente com o avanço dos estudos para classificação de gênero.

Palavras-chave

Classificação de Gênero, Linguagem Subjetiva, Léxicos, Blog.

1. INTRODUÇÃO

No atual contexto da massificação da internet, as pessoas compartilham dados a todo instante na web. Muitos dos dados são de autores anônimos, ou não o são acompanhados de metadados com informação do gênero autoral. Portanto, desenvolvimento de técnicas automatizadas para realizar a Classificação do Gênero Autoral (CGA) pode ser de bastante relevância em diversas aplicações onde os serviços consumidos por homens e mulheres são diferentes (Carloto; 2001).

Empresas podem se beneficiar de informações sobre gênero autoral, com a finalidade de obter uma melhor inteligência de mercado, pois os dados podem ser explorados em publicidade direcionada, por exemplo. Pesquisas em CGA podem auxiliar companhias através da análise dos tópicos mais comentados por homens ou mulheres. Tendo conhecimento sobre esses dados, as empresas podem ver quais produtos que atendem melhor cada um dos públicos, a partir do gênero do autor, podendo prestar um serviço personalizado aos usuários. Além disso, a administração pública também pode se beneficiar, através do melhor direcionamento de campanhas.

Este artigo apresenta os resultados de uma investigação com o objetivo de desenvolver uma abordagem para CGA baseada na subjetividade da linguagem, considerando que o nível de subjetividade nos textos escritos por homens é diferente daquele observado nos textos escritos por mulheres. Ellyn Rolleston Keith (2017) acredita que ao comparar as frases entre homens e mulheres, o sexo feminino tende a ser mais consistente com palavras positivas. Em contrapartida, ela supõe que os homens tendem a escrever mais negativamente. Nossa intenção é utilizar léxicos baseando-se na subjetividade das palavras com a finalidade de encontrar diferenças na escrita entre os gêneros.

Outras pesquisas em classificação de gênero foram realizadas utilizando *Natural Language Processing* (NLP). No entanto, a maioria foi relacionada a textos formais, como notícias. Nossa pesquisa se diferencia visto que utiliza publicações retiradas de blogs, as quais geralmente utilizam-se da linguagem informal, possuindo menor rigor em relação à qualidade da escrita Alguns trabalhos também já pesquisaram sobre a classificação de gênero autoral em blogs, estes realizaram a análise do conteúdo utilizando palavras de relevância, dicionários, *Part of Speech* (POS) e seleção de recursos, juntamente com algoritmos de aprendizagem (Schler, J. et al., 2006; Argamon et al., 2007; Yan e Yan, 2006).

O Dataset utilizado nesta pesquisa é o mesmo do artigo de Mukherjee & Liu (2010), o qual conta com cerca de 3.200 textos retirados dos blogs. A avaliação é realizada com base nas distâncias semânticas entre o conteúdo textual dos blogs e de léxicos denotando dimensões de subjetividades, tais como sentimento, valoração e modalização. Presumimos que, conforme os textos dos blogs se aproximam dos léxicos que selecionamos, mais subjetivo eles são. Desta forma, baseado nessa subjetividade, poderemos verificar se as mulheres escrevem de maneira diferente comparado aos homens. Portanto, neste trabalho, buscamos responder a seguinte questão de pesquisa: *Quão eficaz é a representação da subjetividade para a classificação de gêneros?*

Nós utilizamos léxicos separadamente, buscando encontrar a distância entre eles e os textos dos blogs. Após conseguirmos valores para as distâncias, utilizamos os resultados com algoritmos de aprendizagem buscando resultados de precisão. Ao final do artigo, poderemos responder se os léxicos baseados em subjetividade são uma boa forma de realizar classificação de gênero ou não.

2. TRABALHOS RELACIONADOS

Atualmente, existem alguns trabalhos com foco em classificação de gênero. Estes, por sua vez, utilizaram textos relacionados a blogs como dados de suas pesquisas, a exemplo de Schler et al. (2006); Argamon, et al. (2007); Yan & Yan (2006); Nowson et al. (2005). As técnicas de classificação propostas nestes trabalhos baseiam-se em *Part of Speech* (POS). A utilização de POS é feita empregando algoritmos que associam termos discretos do texto e partes ocultas da fala a um conjunto de tags descritivas. Nas quais essas tags usam algoritmos de marcação aplicando POS com o intuito de se dividir em dois grupos distintos: baseados em regras e os estocásticos. Classes de palavras (e.g., substantivo, adjetivo) ligadas a algoritmos de classificação também foram utilizadas para capturar o sexo dos escritores.

Parte das pesquisas sobre classificação de gênero utiliza o POS como métrica principal para a categorização. Os autores Mukherjee & Liu (2010) também empregam um pouco de POS na sua pesquisa, porém o maior foco do trabalho é saber o quão bem seu algoritmo *Ensemble Feature Selection* (EFS) consegue especificar o sexo do autor. Este algoritmo se baseia em um método de características estilísticas, o qual tem por função capturar o estilo de escrita das pessoas. Os autores lidam com POS relacionando-o com os n-gramas, tais como os utilizados por Koppel et al. (2002) e Argamon et al. (2007) que empregaram como recurso o POS 3-gramas, 2-gramas e unigrama. Os unigramas usados por Mukherjee & Liu (2010) possuem comprimento variável com a função de capturar as regularidades da fala. Os padrões de POS, conforme utilizados por Mukherjee & Liu (2010), podem lidar com n-gramas e buscar regularidades adicionais em sequência, fazendo com que se diferencie das outras pesquisas e discriminando melhor os textos.

Mukherjee & Liu (2010) fizeram uso de *F-measure* (Heylighen & Dewaele, 2002) com base na frequência de uso do POS. Uma pontuação mais baixa no *F-measure* aponta contextualidade, com maiores marcas da utilização de pronomes, verbos, advérbios e interjeições. Por outro lado, uma pontuação mais alta mostra formalidade, apresentada através de substantivos, adjetivos, preposições e artigos. Essa métrica não foi utilizada para classificar o sexo do autor, mas para descobrir a noção de implicitividade do texto, em oposição a formalidade. Continuamente eles buscaram formas de apreender recursos a partir do estilo de escrita, os quais são capturados das seguintes maneiras: POS, palavras de acordo com o contexto do blog e termos que aparecem com maior frequência. Além disso, os escritores utilizam recursos preferenciais por gênero e também análise fatorial de palavra.

Neste artigo pretendemos contribuir com a classificação de gênero de uma forma diferente, através da utilização de léxicos baseados na subjetividade da linguagem. Diferentemente das pesquisas realizadas anteriormente, tivemos a intenção de focar nos aspectos semânticos do conteúdo textual dos blogs, com base em análises de diferentes dimensões de subjetividade. Buscando o melhor resultado para distinguir os gêneros, além da utilização de um léxico que representa vários aspectos de subjetividade, utilizamos ainda outro léxico que representam aspectos de sentimento, por ser uma dimensão da linguagem subjetiva com grande potencial de diferenciação de gênero autoral. Nossos experimentos foram realizados com o dataset disponibilizado por Mukherjee & Liu (2010), que consiste em artigos de blogs de tópicos diversos, escritos por homens e mulheres (em inglês).

3. METODOLOGIA

Existem diferentes maneiras de realizar a classificação de gênero. Entre as mais utilizadas estão análise de POS, de características estilísticas, análise fatorial, entre outras. Nosso método busca utilizar léxicos que representam diferentes dimensões de subjetividade, verificando o quão próximo os textos escritos por autores de ambos os sexos se aproximam de cada um desses léxicos, retornando um valor que representa a distância semântica entre os mesmos.

3.1 Word2Vec

Durante nossa pesquisa utilizamos word2vec, uma técnica desenvolvida por pesquisadores do google, a qual tem por ideia transformar cada palavra (*token*) de um conjunto de frases em um vetor numérico no qual possa se encontrar semântica. Para entender melhor, citamos um exemplo mostrado pelos autores: se pegarmos um token isolado chamado *Madri* e subtrairmos de um vetor representado por *Espanha* e então logo após adicionarmos um token cuja sua representação é *França*, o vetor resultado muito provavelmente será *Paris* ou um valor muito próximo daquele associado a esta palavra. Podemos representar o que foi descrito anteriormente na Equação 1:

$$\text{vec("Madri")} - \text{vec("Espanha")} + \text{vec("França)} \approx \text{vec("Paris")} \quad (1)$$

Word2Vec (Rong, Xin. 2014) faz parte de uma classe de modelos denominada de linguagem neural *Neural Language Models* (NLM), visto que o mesmo emprega redes neurais profundas (*Deep Learning - DL)* para realizar o aprendizado. O treinamento do Word2Vec é realizado a partir de dois algoritmos: *Continuous Bag of Words* (CBOW) e *Skip-gram* (SG). O CBOW apresenta a ideia de prever a próxima palavra que estaremos buscando dentro de um determinado conjunto de palavras. Para tal feito, ele utiliza uma rede neural que recebe como entrada um vetor denominado *onehot encoded*, que representam palavras do contexto; e, como saída, retorna a palavra que estava-se esperando. O SG, por sua vez, funciona de maneira inversa. Utilizando como ponto inicial uma palavra qualquer, o objetivo do algoritmo é encontrar o contexto do qual essa palavra foi originada. Da mesma forma do CBOW, utiliza-se uma rede neural, onde recebe-se um vetor que represente a palavra que procuramos e retorne um conjunto de palavras, o qual este associado com a primeira palavra buscada.

Além do que foi apresentado, é necessário que exista uma correspondência entre os vetores. Uma das formas aplicadas para realizar essa semelhança é através da similaridade de cossenos. Temos que o produto entre dois vetores é uma operação algébrica que o transforma em um produto escalar. Considerando dois vetores separadamente *a* e *b* com a mesma dimensão, o produto entre *a* e *b* é apresentado na Equação 2:

$$< a, b > \; = \; ||a||\, x\, ||b||\, x\, cos(a, b) \quad (2)$$

A partir da fórmula acima, podemos isolar o *cos*(*a*,*b*) e teremos a Equação 3:

$$cos(a, b) \; = \; \frac{<a,b>}{||a||\, X\, ||b||} \quad (3)$$

Aprendemos que o valor do cosseno pode variar entre -1 e 1, entre dois ângulos. Quando o ângulo chega próximo a 0º o valor do cosseno se aproxima de 1; inversamente, quando o ângulo se aproxima de 180º o resultado do cosseno aproxima se de -1; em 90º, o cosseno é 0. Portanto, podemos entender que, quando o cosseno está próximo de 1 e temos um ângulo mais agudo, significa que temos uma grande similaridade de vetores. Diferentemente, se o ângulo for obtuso, temos o resultado oposto, mostrando oposição entre os vetores. Então, percebemos que a afinidade entre os cossenos é obtida a partir do cálculo do cosseno, o qual acontece entre os ângulos dos vetores que temos por objetivo comparar utilizando a fórmula do produto entre dois vetores.

3.2 Word Mover's Distance (WMD)

Após utilizarmos o Word2Vec, utilizamos o Word Mover's Distance (WMD) (Huang, Gao, et al. 2016), o qual se baseia na integração de palavras que aprendem com a representação semântica através da co-ocorrência em locais diferentes das frases. O WMD utiliza o resultado de técnicas avançadas como a Glove (Pennington et. al., 2014) e o já comentado Word2Vec, com a finalidade de gerar palavras de qualidade sem precedentes e escalar para grandes conjunto de dados. Elas também demonstram

que os relacionamentos semânticos geralmente são preservados nas operações de vetores de palavras. Traduzindo literalmente, Word Mover's Distance significa Distância do Motor das Palavras, o qual demonstra que as distâncias entre os vetores das palavras; apresentam, em algum nível, uma semântica significativa. O algoritmo utiliza essa propriedade de integração de vetores para tratar os textos dos blogs como se fosse uma nuvem de pontos com várias palavras dentro desta nuvem. .

Supondo que tomamos dois textos separadamente e chamamos de A e B, a distância entre os dois textos é calculada através da cumulativa do intervalo em que as palavras do texto A precisam percorrer para corresponder exatamente com as a nuvem de pontos que está no texto B. O WMD também tem por objetivo abordar a sintática e a semântica das palavras para resultar uma melhor semelhança entre documentos do texto. O algoritmo mede a diferença entre dois textos com o propósito de encontrar o "caminho" mais curto entre as palavras de um texto e outro. Calcular as distâncias através do WMD é interessante, pois é simples de entender, é interpretável e utiliza conhecimentos como Word2Vec e Glove.

Vale a pena utilizar o WMD pois ele nos proporciona baixas taxas de erros perante o conjunto de dados. Conforme a quantidade de palavras vai sendo ampliada e mais dados vão sendo inseridos para o treinamento, obtém-se gradativamente um melhor resultado.

Na Figura 1 é apresentado um exemplo típico de WMD:

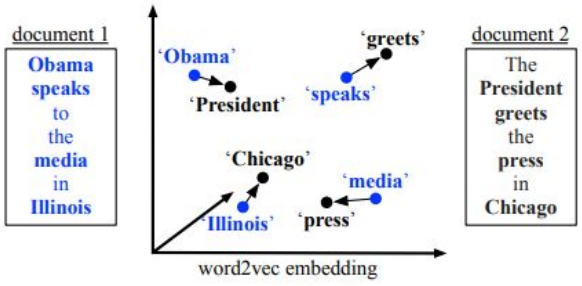

Figura 1. Imagem retirada da publicação de Kusner et al [8] .

Exceto pelas *StopWords*[1], não existem palavras comuns entre as duas sentenças. Porém, as duas abordam o mesmo tópico. Através de WMD, calcula-se o custo mínimo para transportar as palavras da sentença 1 para a sentença 2. Neste caso, palavras semanticamente próximas como “obama” e “presidente” incorrerão em um menor custo de transporte, denotando maior similaridade entre as sentenças.

3.3 Léxicos

A personalidade diferencia uma pessoa de outra. Sabemos que cada indivíduo possui uma maneira específica de agir em cada situação. Acreditamos que isso também pode ser refletido no momento da escrita. Além disso, da mesma forma que homens e mulheres, em geral, apresentam personalidades distintas, existe a possibilidade dos mesmos se diferenciarem também na maneira de escrever. Pesquisas na área de psicologia sugerem que alguns traços de personalidade possuem correlação com o comportamento linguístico (Homes J, 2006; Henley et. al. 1991). A hipótese que guia essa investigação é que existe diferença no uso de linguagem subjetiva entre homens e mulheres.

Nós utilizamos 12 léxicos de subjetividade em inglês, criados por linguistas da área, descritos abaixo:

- Verbos assertivos. São aqueles cujas cláusulas de complemento afirmam uma proposta (e.g., *appear*, *predict*, *certain*). Proposto por Joan B. Hooper (1975).
- Verbos factivos. São aqueles que pressupõem a verdade da cláusula de complemento (e.g., *resent*, *forget*, *remember*). Proposto por Kiparsky and Kiparsky (1970).
- Verbos limitantes, são usados para reduzir o compromisso com a verdade através de um proposição, evitando previsões ousadas (e.g., *frequently*, *lagerly*, *could*). Proposto por Ken Hyland (2015)..
- Verbos implicativos. São aqueles que a partir de uma sentença principal afirmada com um desses verbos como predicado compromete o falante a uma proposição implícita (e.g., *manage*, *succeed*, *prevent*). Proposto por Lauri karttunen (1971)..
- Palavras de positividade baseadas na fala emocional (e.g., *vivacious*, *delicious*, *incredible*). Proposto por Hu & Liu (2004.).
- Palavras de negatividade baseado na fala emocional (e.g., *decadence*, *horrendous*, *afraid*). Proposto por Hu & Liu (2004).
- Palavras de subjetividade, são aquelas que caracterizam uma opinião ou atitude marcada por sentimentos (e.g., *article*, *collapse*, *nationalist*). Proposto por Riloff, E. & Wiebe, J. (2003).
- Verbos de relatório. São aqueles que mantém sempre a verdade de um palavra ou frase. Por exemplo, a palavra "assassinato" implica em matar, porque não há como matar sem assassinar (e.g., *caution*, *illustrate*, *negotiate*). Proposto por Recasens M., Danesco C,. & Jurafsky D. (2013).
- Palavras de efeito positivo. Léxicos que foram filtrados por máquina, significando maior imprecisão (e.g., *induce*, *commend*, *step*). Proposto por Choi & Wiebe (2014).
- Léxicos baseados nas palavras com efeito positivo, propostos por Choi & Wiebe (2014). Estes léxicos são os mesmos do anterior, porém com uma melhor precisão. As palavras foram filtradas por pessoas e não pela máquina (e.g., *recognize*, *revolutionise*, *admire*).
- Palavras de efeito negativo. Léxicos que foram filtrados pela máquina, significando maior imprecisão (e.g., *trash*, *argufy*, *scrimp*). Propostos por Choi & Wiebe (2014).
- Léxicos baseados nas palavras com efeito negativo, também propostos por Choi & Wiebe (2014). Estes léxicos são os mesmos do anterior, porém com uma melhor precisão. As palavras foram filtradas por pessoas e não pela máquina (e.g., *lose*, *divide*, *bungle*).

Nossa abordagem para mensurar subjetividade linguística se baseia no cálculo da distância semântica de um texto de blog para cada um dos léxicos apresentados acima. Isso resulta em um vetor de 12 dimensões, onde cada dimensão representa a distância entre o texto de destino e um léxico. Instintivamente, cada dimensão quantifica os valores acerca de certa linguagem subjetiva dentro dos blogs alvos, retornando para nós esses valores que serão utilizados juntamente com os algoritmos de aprendizagem.

3.4 Sentenças

Decidimos também utilizar uma forma diferente para medir as distâncias entre as palavras dos textos e os léxicos. Utilizamos as técnicas apresentadas acima, porém, com uma diferença.

[1] *Stop Words* são palavras que podem ser consideradas irrelevantes em diversos problemas na área de recuperação de informação, tais como em motores de busca. São exemplos de *stop words*: a, as, os, e, os, de, para, com, sem.

Decidimos dividir os textos retirados dos blogs em sentenças (frases). No caso, a cada vírgula, ponto-e-vírgula, ou ponto final que nosso algoritmo encontra, ele define aquela parte do texto em uma sentença. Utilizando os resultados do WMD, atribui-se um valor a cada sentença, representando sua distância semântica a um determinado léxico. Em seguida, o algoritmo realiza a média total dos valores das sentenças. Em conjunto a esse método, utilizamos dois léxicos diferentes especializados em sentimentos, baseados em sentimento, léxicos foram propostos por Nielson et al. (2011) e foram utilizados originalmente para análise de sentimentos com dados do twitter. Resolvemos utilizar também esses léxicos com base na hipótese de que sentimentos podem ser uma dimensão da subjetividade com grande potencial discriminativo entre as linguagens empregadas por homens e mulheres.

3.5 Dataset

Com a função de tentar manter o problema da classificação de gênero o mais geral possível, os dados foram coletados de muitos sites que hospedam blogs, como, por exemplo, *bloger.com*, *technorati.com,* entre outros. Utilizamos o dataset publicado por Mukherjee & Liu (2010), o qual possui um conjunto de dados de 3.226 textos retirados dos blogs, de maneira que cada texto já está rotulado com o sexo do autor. De acordo com os autores, o sexo dos autores dos textos foram determinados visitando o perfil dos mesmos. Imagens e avatares também foram utilizadas para ajudar na confirmação do gênero. No final, para garantir a rotulagem dos gêneros, dois grupos de estudantes foram utilizados, um deles para etiquetar os textos e outro grupo para a conferência. Dos 3.226 textos, 1.678 (cerca de 52%) foram escritos por homens e 1.548 (cerca de 48%) foram escrito por mulheres. A média de tamanho para cada texto é de 250 palavras para homens e 320 palavras para as mulheres. A distribuição dos dados é apresentada na Figura 2.

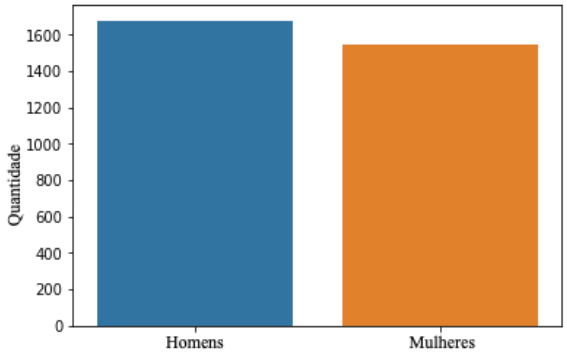

Figura 2. Gráfico mostrando a proporção entre os textos escritos por autores do sexo masculino e feminino.

4. RESULTADOS

Nesta seção iremos apresentar os resultados obtidos a partir das técnicas propostas na Seção 3 e verificar como elas afetam a precisão de classificação. Utilizamos *Random Forest*, *Support Vector Machine* e *XG Boost* como algoritmos de aprendizagem. Em todos os resultados experimentais, utilizamos o *recall*, *precision* e *F-measure* como fator de precisão.

4.1 Resultados complementares

Após realizarmos os procedimentos ditos acima, os resultados gerados foram médias que variam entre 0 e 1, no qual um valor mais próximo do 0 significa menos proximidade das palavras dos blogs com os léxicos escolhidos pelos linguistas. No momento em que rodamos os algoritmos, passamos as palavras para medir suas distâncias com cada léxico. Em seguida, geramos gráficos box-plots com os resultados a fim de visualizar bem as diferenças entre as escritas de homens e mulheres. O box-plot é a representação gráfica de um conjunto de dados, o qual permite avaliar a dispersão dos mesmos, destacando também os valores discrepantes dos conhecimentos comparados.

Nos gráficos apresentados a seguir, a letra F denota os autores do sexo Feminino, enquanto a letra M representa os autores do sexo Masculino. Nestas representações gráficas, quanto menor os valores apresentados, menores são as distâncias aos léxicos analisados, ou seja, mais subjetivos os textos são. Na Figura 3 são mostrados os gráficos box-plots com os resultados obtidos a partir dos léxicos. Os léxicos baseados em palavras de positividade alcançaram os melhores resultados dentre os estudados. Ao lado esquerdo temos os gráficos dos verbos assertivos e ao direito os verbos factivos. Podemos perceber que os resultados foram muito próximos. Comparando primeiramente apenas os box-plots dos verbos assertivos, notamos que a distinção entre os valores de subjetividade das mulheres é imperceptível em relação aos homens. Adicionalmente, vemos que os resultados dos verbos factivos também são muito próximos, notando-se que os quartis de ambos os léxicos baseado nos verbos estão bem juntos, dificultando, dessa forma, a classificação do gênero. Resulta-se que não podemos afirmar qual o sexo do escritor através desses resultados.

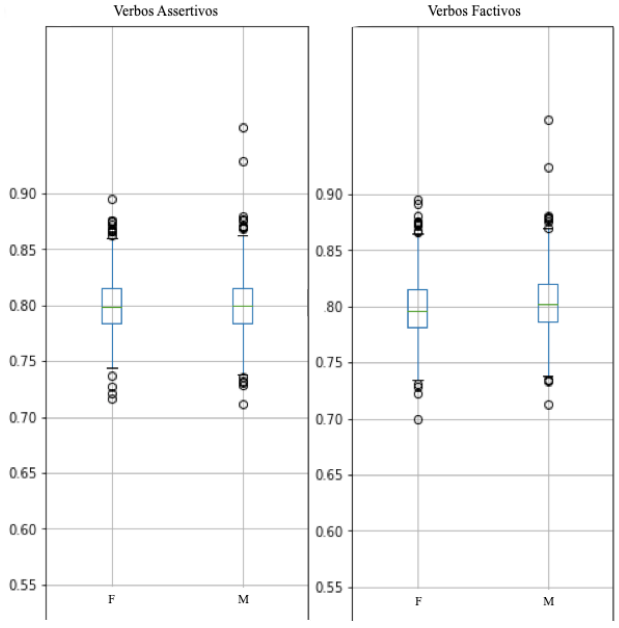

Figura 3. Distâncias semânticas obtidas entre os textos de autores do sexo Feminino (F) e Masculino (M) à léxicos representando verbos assertivos e factivos, respectivamente.

Na Figura 4, observamos os verbos limitantes ao lado esquerdo e os verbos de relatório ao lado direito. Buscando fundamentação nestes léxicos, percebemos que os dois possuem mediana e os dois quartis entre os valores de 0.75 e 0.80. A partir desses resultados, percebemos que o segundo quartil dos verbos limitantes tem um valor maior, comparado ao mesmo quartil dos verbos de relatório. Isso nos mostra que tanto os homens quanto as mulheres tendem a usar um pouco mais dos verbos limitantes nas suas escritas. Não apenas o segundo quartil como também a mediana está um pouco mais elevada. Mesmo contendo essa sutil diferença, os resultados não são promissores, uma vez que a discrepância entre a escrita dos homens e das mulheres é quase inexistente para esses léxicos.

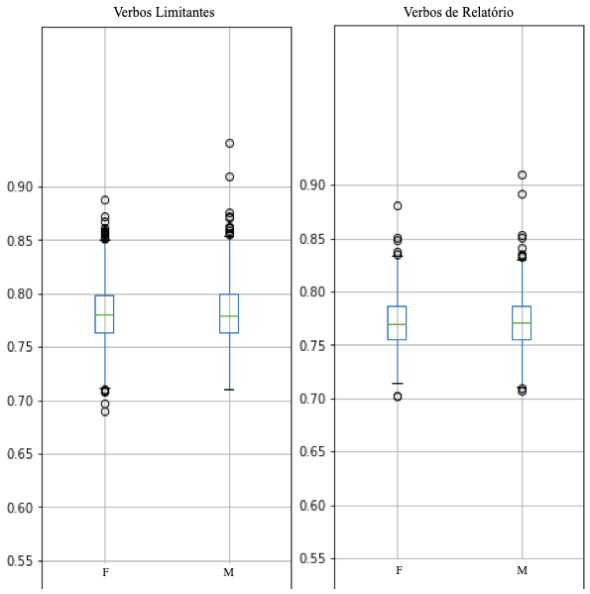

Figura 4. Distâncias semânticas obtidas entre os textos de autores do sexo Feminino (F) e Masculino (M) à léxicos representando verbos limitantes e de relatório, respectivamente.

Na Figura 5, temos os léxicos de palavras com efeito positivo ao lado esquerdo e de palavras com efeito negativo ao lado direito. Como já foi relatado, as palavras desses léxicos foram retiradas de outro léxico, porém com a avaliação de pessoas. Com esses léxicos, esperávamos melhores resultados, porém não o aconteceu. Percebe-se claramente que os valores, tanto para palavras positivas quanto para as palavras negativas, foram muito próximos, mostrando que é difícil gerar uma disparidade nas escritas de gêneros. As medianas de ambos os gráficos ficaram entre os valores de 0.75 e 0.80, assim como os verbos limitantes e de relatório. O resultado é que a desigualdade entre os box-plots é bem pequena, impedindo uma boa classificação realizada pela máquina.

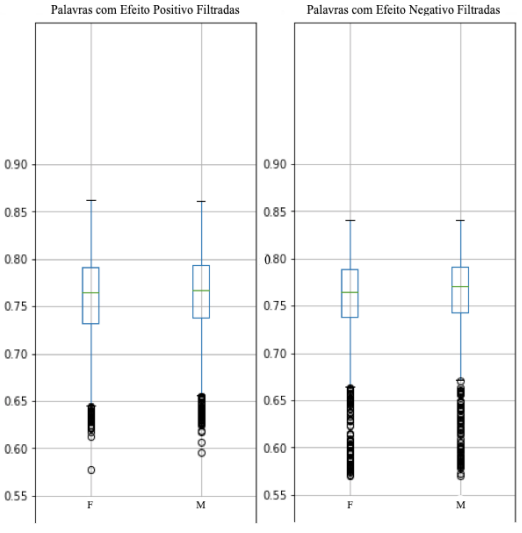

Figura 5. Distâncias semânticas obtidas entre os textos de autores do sexo Feminino (F) e Masculino (M) à léxicos representando palavras com efeito positivo e negativo, respectivamente. Ambos filtrados por humanos..

Na Figura 6, apresentam-se quatro box-plots. O primeiro deles representa os resultados dos léxicos com base nos verbos implicativos. O segundo representa os resultados de palavras subjetivas. O terceiro, palavras com efeito negativo. O quarto, por sua vez, palavras negativas. Notamos que, dos quatro gráficos, nenhum deles pode nos resultar em uma boa discrepância entre os valores das escritas de homens e mulheres. O resultado dos léxicos baseado em verbos implicativos, são os melhores em comparação aos quatro, pois ele é o único que tem a mediana do box-plot masculino, ligeiramente acima da mediana feminina. Os demais resultados possuem valores muito próximos, que não ajudam o algoritmo a ter boas saídas para classificação. Percebe-se então que cada vez mais, usando diferentes palavras e diferentes verbos para comparar as escritas de gêneros, fica difícil obter um boas diferenças..

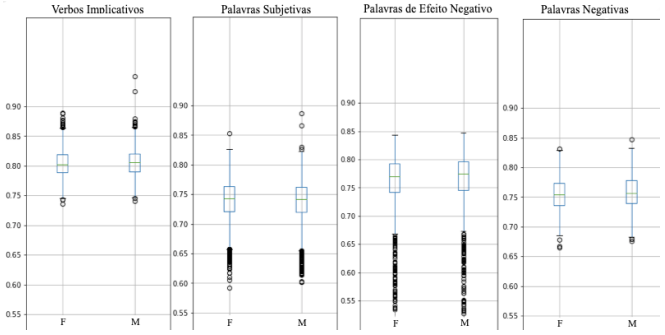

Figura 6. Distâncias semânticas obtidas entre os textos de autores do sexo Feminino (F) e Masculino (M) à léxicos representando verbos implicativos, palavras subjetivas, palavras com efeito negativo e palavras negativas, respectivamente.

Na Figura 7, observamos os melhores resultados, muito embora não fosse o resultado esperado. Primeiramente notamos que as palavras com efeito positivo e palavras positivas, apesar de estarem praticamente com o mesmo nome, são léxicos diferentes. Observando os gráficos, constata-se que a mediana de ambos os box-plot masculino estão um pouco acima em comparação com a das mulheres. Isso nos leva a ver que os homens geralmente tendem a escrever de maneira mais positiva que as mulheres, apesar dos valores ainda serem bem próximos. Mesmo com esses resultados, ainda foi bem difícil para os classificadores determinarem por qual gênero foi escrito algum texto.

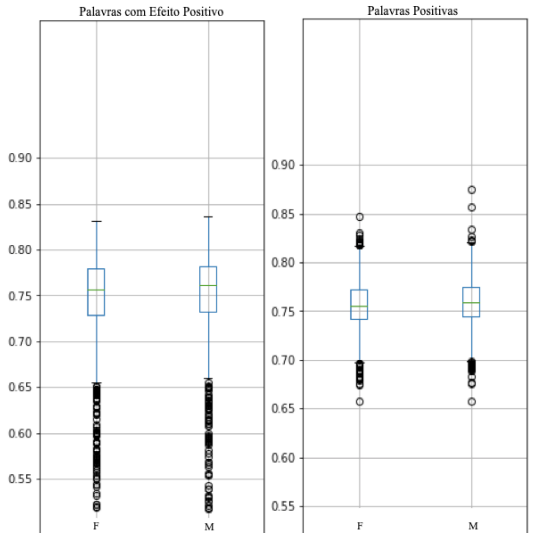

Figura 7. Distâncias semânticas obtidas entre os textos de autores do sexo Feminino (F) e Masculino (M) à léxicos representando palavras com efeito positivo e palavras positivas..

Portanto, a partir dos resultados observados acima, podemos começar a entender que tentar classificar as diferenças de escrita entre gêneros não é algo simples, já que os dois possuem uma forma de escrita muito próxima. Mesmo lidando com vários verbos e palavras distintas com objetivo de encontrar uma boa

disparidade entre os valores, não obtivemos bons resultados. A partir de todos os valores obtidos, utilizamos algoritmos de inteligência artificial, os quais utilizam aprendizagem supervisionada, com a finalidade de treiná-los e obtermos uma noção mais precisa da qual estes dados podem nos mostrar.

4.2 Comparando os algoritmos de aprendizagem supervisionada

Utilizando os dados apresentados na Seção 4.1, executamos três algoritmos de classificação diferentes: *Random Forest*, *Support Vector Machines* (SVM) e *XG Boost*. Nosso objetivo era conseguir visualizar se nosso método obtinha bons resultados para classificar gêneros autorais. Embora os *bloxplots* já demonstrem que as diferenças são muito pequenas, em trabalhos publicados na literatura utilizando técnicas similares para classificação, os resultados também são muito próximos, difíceis de serem percebidos por humanos, mas ao serem utilizados em classificadores, bons resultados são observados.

Entretanto, não obtivemos bons resultados com os classificadores avaliados. Todavia, nosso trabalho tem por inovação a tentativa de utilizar os léxicos para classificação de gêneros, tendo em vista que a maioria dos outros artigos que buscavam classificar gêneros focaram em POS. Começamos testando o *Random Forest*, que cria uma "floresta" de combinações, em que cada árvore de decisão se combina para obter uma predição com maior acurácia e precisão. O *XG-Boost* também é um algoritmo baseado em árvores, porém com a diferença de que ele utiliza uma estrutura chamada de *Gradient Boosting*. Para dados tabulares, esses algoritmos de árvores tendem a apresentar bons resultados (sem generalizar). E o terceiro algoritmo testado foi o *Support Vector Machine* (SVM). Este foi o algoritmo que apresentou melhores resultados dentre os três testados.

Para efetuar os testes, primeiramente dividimos os dados aleatoriamente. Logo após inserimos um conjunto de testes que representa 20% do dataset original. O restante dos dados formaram então os dados de treinamento. Utilizamos as sementes de aleatoriedade com o valor 42. O *Random Forest* obteve o resultado com menor precisão em relação aos três algoritmos. Passamos o hiperparâmetro de estimadores com o valor de 750, o qual indica o número de árvores construída pelo algoritmo antes de tomar uma votação ou fazer uma média. A partir daí utilizamos a biblioteca *Scikit-learn* para efetuar o resto do algoritmo. Já o *XG-Boost* obteve o segundo lugar na nossa comparação. Quando criamos os modelos de gradiente com o *XG-Boost* utilizando o *Scikit-learn*, o parâmetro usado para taxa de aprendizagem pode ser definido com a função de controlar o peso das novas árvores adicionadas ao modelo. Então utilizamos o valor de 0.1 para essa taxa de aprendizagem. E, por último, o algoritmo de aprendizagem que apresentou o melhor resultado foi o *Support Vector Machine* (SVM). Para a realização dos testes com o SVM, utilizamos os próprios parâmetros fornecidos pela biblioteca do *Scikit-learn.*

A Tabela 1 apresenta os valores obtidos com suas respectivas porcentagens, significando que o classificador consegue fazer as predições corretas com essa taxa de assertividade.

	Random-Forest	XG-Boost	SVM
Recall Average	54.30%	57.94%	59.60%
Precision Average	53.59%	56.81%	58.06%
F1-measure Average	53.94%	57.37%	58.82%

Tabela 1. Resultados dos algoritmos de aprendizagem

O *recall* é a fração da quantidade local de instâncias relevantes que foram realmente recuperados. Já o *precision* é a razão entre as observações previstas corretamente e o total de observações previstas positivamente. O *F1-measure* é a média ponderada de precisão e recuperação. Portanto, essa pontuação leva em consideração tanto os falsos positivos quanto os falsos negativos.

4.3 Resultados das sentenças

Nesta subseção iremos retratar os resultados retornados a partir da divisão dos textos dos blogs em sentenças. Como foi explicado na metodologia, nós dividimos cada texto de blogs em frases separadas e o WMD retornava valores para cada sentença dessa e no final retornava um valor para o texto baseado na média das sentenças. O resultado não diverge praticamente nada quando colocado lado a lado com os outros 12 léxicos. Como os léxicos utilizados foram de sentimentos, esperávamos resultados mais discrepantes em relação aos léxicos baseados em palavras positivas. Abaixo, na Figura 6, encontramos os box-plots com os dois léxicos aplicados aos textos dos blogs.

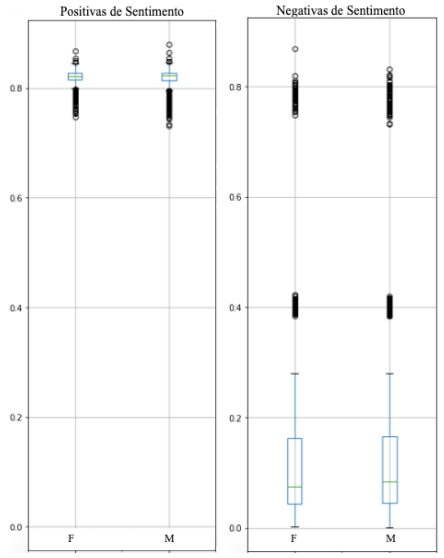

Figura 8. Distâncias semânticas obtidas entre os textos de autores do sexo Feminino (F) e Masculino (M) baseadas nos léxicos escritos por Finn Årup Nielsen

O que podemos observar a partir dos resultados exibidos na na Figura 8 é que, mesmo com léxicos baseados em sentimentos, não é possível encontrar diferenças significativas. Notamos que, dessa vez, houve uma disparidade grande na quantidade de palavras positivas que ambos os sexos usam e nas negativas, as quais foram encontradas poucas vezes ao percorrer os textos. Mesmo existindo essa diferença, os homens e as mulheres continuam escrevendo de maneira muito parecida. Nos gráficos notamos que os box-plots para ambos os léxicos são muito próximos se tratando das escritas, o que dificulta e muito o trabalho da máquina na hora de classificar. A Tabela 2 apresenta os resultados dos algoritmos de aprendizagem, *Random Forest* e *Support Vector Machine* (SVM).

	Random-Forest	SVM
Recall Average	50%	84.82%
Precision Average	54.01%	52.87%
F1-measure Average	51.93%	65.14%

Tabela 2. Resultado dos algoritmos de aprendizagem para os léxicos do Finn Årup Nielsen

Na Tabela 2, podemos observar que os resultados foram muito próximos dos anteriores, porém com a diferença de que o SVM no recall conseguiu uma valor expressivo de 84.82%, muito provavelmente pelo fato de que o recall ter base na relevância dos textos, havendo a probabilidade de ser recuperado em uma pesquisa. O F1-measure também teve um resultado mais significativo quando dividimos os textos em sentenças, chegando a 65.14%.

5. CONCLUSÃO

Este artigo apresentou uma abordagem para CGA baseada em técnicas modernas de processamento de linguagem natural, utilizando aprendizagem profunda, que teve como premissa a diferenciação de gêneros com base no emprego de linguagem subjetiva. Embora a literatura apresente outros trabalhos em CGA, nenhum deles fez uso de abordagem baseada na subjetividade linguística. Nós utilizamos distâncias semânticas a léxicos especializados para mensurar o nível de subjetividade dos textos produzidos por homens e mulheres. Conduzimos experimentos de CGA utilizando diferentes algoritmos de aprendizagem de máquina que já foram aplicados com sucesso em problemas similares, porém, não foi possível obter bons resultados de classificação. Estes achados demonstram que a CGA é uma tarefa que ainda está em aberta e que pode se beneficiar de outras investigações futuras.

Para trabalhos futuros, pretende-se aplicar técnicas diferentes do WMD para o cálculo da distância semântica, bem como utilizar outros léxicos reportados na literatura de maneira que representem outros aspectos da linguagem que tenham potencial significativo de distinguir homens de mulheres, à luz das investigações reportadas na literatura da área de psicologia.

6. REFERÊNCIAS

Argamon, Shlomo, et al. "*Gender, genre, and writing style in formal written texts.*" Text-The Hague Then Amsterdam Then Berlin- 23.3 (2003): 321-346.

Argamon, Shlomo, et al. "*Mining the blogosphere: Age, gender and the varieties of self-expression*." First Monday 12.9 (2007).

Arjun Mukherjee, and Bing Liu,. *Improving Gender Classification of Blog Authors,* In Proceedings of the 2010 conference on Empirical.

Carloto, Cassia Maria. "*O conceito de gênero e sua importância para a análise das relações sociais.*" Serviço Social em Revista, Londrina 3.2 (2001): 201-213.

Nielsen, F. Å. (2011). A new ANEW: *Evaluation of a word list for sentiment analysis in microblogs*. arXiv preprint arXiv:1103.2903.

Henley, Nancy, and Cheris Kramarae. "*Miscommunication, gender, and power*." Miscommunication and Problematic Talk (1991): 18-43.

Heylighen, F., and Dewaele, J. 2002. *Variation in the contextuality of language*: an empirical measure. Foundations of Science, 7, 293–340.

Holmes, Janet. *Women, men and politeness*. Routledge, 2013.

Hu, M., & Liu, B. (2004, August). *Mining and summarizing customer reviews*. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM.

Huang, Gao, et al. "*Supervised word mover's distance.*" Advances in Neural Information Processing Systems. 2016.

Hyland, K. (2018). *Metadiscourse: Exploring interaction in writing*. Bloomsbury Publishing.

Keith, Ellyn Rolleston. "*A Sentiment Analysis of Language & Gender Using Word Embedding Models.*" (2017).

Koppel, M., Argamon, S., Shimoni, A. R.. 2002. *Automatically Categorizing Written Text by Author Gender.* Literary and Linguistic Computing.

Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015, June). *From word embeddings to document distances.* In International conference on machine learning (pp. 957-966).

Nowson, S., Oberlander J., Gill, A. J., 2005. *Gender, Genres, and Individual Differences.* In proceedings of the 27th annual meeting of the Cognitive Science Society (p. 1666- 1671). Stresa, Italy.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. "*Glove: Global vectors for word representation*." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.

Recasens, M., Danescu-Niculescu-Mizil, C., & Jurafsky, D. (2013, August). *Linguistic models for analyzing and detecting biased language*. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1650-1659).

Rong, Xin. "*word2vec parameter learning explained.*" arXiv preprint arXiv:1411.2738 (2014).

Schler, J., Koppel, M., Argamon, S, and Pennebaker J. 2006. *Effects of age and gender on blogging*, In Proc. of the AAAI Spring Symposium Computational Approaches to Analyzing Weblogs.

Srilakshmi Bharadwaj, Srinidhi Sridhar, Rahul Choudhary, Ramamoorthy Srinath, "*Persona Traits Identification based on Myers-Briggs Type Indicator(MBTI) - A Text Classification Approach*", Advances in Computing Communications and Informatics (ICACCI) 2018 International Conference on, pp. 1076-1082, 2018.

Word2vec. Disponível em : <https://code.google.com/archive/p/word2vec/> . Último acesso em: 24 out. 2019.

Yan, Xiang, and Ling Yan. "*Gender Classification of Weblog Authors.*" AAAI spring symposium: computational approaches to analyzing weblogs. 2006.