Inferindo o sexo de usuários de redes sociais utilizando o LIWC em português do Brasil

Luiz Antonio da Ponte Junior [1], **Gustavo Paiva Guedes**[1],**Eduardo Bezerra**[1]

[1]CEFET/RJ - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca
Av. Maracanã, 229 - Rio de Janeiro - RJ - Brasil.

lapjunior@yahoo.com.br,{gustavo.guedes,eduardo.silva}@cefet-rj.br

Abstract. *This work presents preliminary results of an experimental evaluation about the inference of users gender in a Brazilian social network. This is done using a feature extraction process associated with those users. To achieve this goal, we use a Portuguese version of the linguistic feature called LIWC. Initial experimental results allow us to conclude that the classification task using datasets produced with LIWC is able to present satisfactory results. This occurs with no direct influence of words that manifest gender traits.*

Resumo. *Esse trabalho apresenta resultados preliminares de uma avaliação experimental sobre a possibilidade de se inferir o sexo dos usuários de uma rede social brasileira. Isso é feito a partir de um processo de extração de características associadas a esses usuários. Para alcançar esse objetivo, utilizamos uma versão em português do recurso linguístico denominado LIWC. Os resultado experimentais iniciais permitem concluir que a tarefa de classificação sobre os conjuntos de dados produzidos com o LIWC é capaz de apresentar resultados satisfatórios. Isso é feito sem influência direta das palavras que possuem traços de gênero.*

1. Introdução

O estudo das características de usuários de redes sociais online vem se expandindo nos últimos anos. Essa expansão se deve, em grande parte, ao entendimento de que os resultados desses estudos podem ser aplicáveis em áreas diversas como Marketing, Sociologia, Computação, e Linguística. Diversas redes sociais populares (e.g. Twitter, Facebook) têm tido seus dados coletados para serem estudados por pesquisadores de diversas áreas. Esses estudos adotam diferentes representações dos usuários, seja a partir de seus dados pessoais (e.g., idade, religião), regionais (e.g. local de nascimento, local de trabalho) ou mesmo a partir dos termos presentes em suas postagens. Alguns estudos indicam que é possível determinar os estados emocionais de um indivíduo a partir das palavras por ele utilizadas [Pennebaker 2013].

O objetivo deste trabalho é realizar uma avaliação experimental acerca da possibilidade de se inferir o sexo dos usuários de rede sociais a partir dos aspectos emocionais presentes em seus textos. Para isso, utilizamos o recurso linguístico denominado LIWC, que permite calcular o grau de uso de diferentes categorias de palavras. Esse recurso foi proposto por [Pennebaker et al. 2001]. Em particular, utilizamos neste trabalho a versão em português do LIWC [Filho et al. 2013].Na avaliação experimental, utilizamos dados de usuários de uma rede social online brasileira denominada *Meu Querido Diário* (MQD).

O restante desse trabalho está organizado como segue. Na Seção 2, são apresentados os trabalhos relacionados. Na Seção 3, após uma descrição introdutória acerca do LIWC e do MQD, apresentamos o processo de extração de características dos usuários. Na Seção 4 são discutidos os resultados experimentais. Por fim, a Seção 5 expõe as conclusões e cenários futuros de expansão deste trabalho.

2. Trabalhos Relacionados

Existem trabalhos similares ao apresentado neste artigo que atendem à língua inglesa (e.g. [Peersman et al. 2011, Kokkos and Tzouramanis 2014]). Também podemos observar a existência de estudos capazes de inferir a idade e o sexo em textos em português do Brasil [Araújo et al. 2014]. Entretanto, nesse último, não foram considerados aspectos emocionais, cognitivos ou estruturais dos textos, o que caracteriza uma diferença e uma contribuição do presente trabalho.

O trabalho realizado em [Schler et al. 2006] identifica que o conteúdo dos textos escritos por homens e mulheres possui diferenças. Esse trabalho indica que essas diferenças podem ser identificadas não apenas por temas como política, dinheiro e vida pessoal, mas também pelos estilos de escrita, como o uso de pronomes, artigos e preposições. Essas diferenças presentes no conteúdo dos textos não delimitam apenas o sexo do usuário, mas também são observadas de acordo com faixas etárias.

O estudo apresentado em [Goswami et al. 2009] indica que também existe relação entre o tamanho médio dos textos, o tamanho médio de cada palavra de um texto e o uso de gírias com a faixa etária de grupos de usuários. Por meio de um conjunto montado para treinamento e do uso de um método de classificação, foi possível inferir a faixa etária com uma precisão de 90%. Foi identificado que, em grupos de usuários jovens, é predominante o uso de gírias e mensagens mais curtas.

O trabalho de [Filho et al. 2014] visa inferir o sexo e a idade por meio de textos de usuários coletados do Twitter. O procedimento consistiu na coleta dos últimos 200 textos de cada usuário, no pré-processamento dos textos e na classificação da relevância dos termos utilizados pelos usuários. Foram empregados 4 diferentes algoritmos de classificação (Naive Bayes, Naive Bayes Multinomial, SVM e Random Forest). Os autores utilizaram a medida de ganho de informação para selecionar os atributos mais relevantes. Os métodos propostos alcançaram bons resultados.

O estudo produzido em [Nguyen et al. 2013] destina-se a determinar a relação entre o uso da linguagem e a idade de usuários do Twitter. O trabalho analisou contas holandesas e constatou que entre o comportamento de pessoas jovens havia o maior uso da primeira pessoa e o alongamento de palavras. Por outro lado, o comportamento observado em pessoas mais velhas foi o maior uso de preposições e mensagens mais longas.

3. Inferência do sexo em usuários de redes sociais utilizando o LIWC

O objetivo dessa seção é descrever os procedimentos adotados para avaliar a adequação do dicionário da língua portuguesa do LIWC para inferir o sexo em usuários de redes sociais. Na Seção 3.1 apresentamos uma breve descrição do LIWC, em seguida, na Seção 3.2 detalhamos as características do conjuto de dados `MQD1016`. Na Seção 3.3 descrevemos o processo utilizado na extração de características dos usuários.

3.1. LIWC

O LIWC (*Linguistic Inquiry and Word Count*) é uma ferramenta proposta em [Pennebaker et al. 2001] e tem o objetivo de analisar os componentes emocionais, cognitivos e estruturais de textos. Isso é feito com base em um dicionário de palavras, que pode ser encontrado em diversas línguas. Nesse trabalho utilizamos o dicionário do LIWC para o português do Brasil, encontrado em [Filho et al. 2013].

O dicionário do LIWC em português possui 127.149 palavras, em que cada uma pode ser assinalada a uma ou mais categorias. Essas categorias representam perspectivas linguísticas, psicológicas, dentre outras. O dicionário é composto por um total de 64 categorias (e.g. posemo, pronoum, time). Com isso, cada palavra presente no texto dos usuários pode ser expressa por diferentes perspectivas (categorias).

Diversos estudos existentes utilizam a versão inglesa do LIWC, dentre eles pode-se destacar [Golbeck et al. 2011] e [Schwartz et al. 2013]. O primeiro tem o objetivo de auxiliar na inferência da personalidade de usuários do twitter. O último tem estuda a correlação entre comportamento e personalidade. Embora haja diversos trabalhos com a versão inglesa do LIWC, pouco se encontra sobre o LIWC em português.

3.2. A Rede Social MQD

A rede social MQD (`http://www.meuqueridodiario.com.br`) foi desenvolvida em 2009 e funciona como um diário online. Seus usuários escrevem entradas registrando o que fizeram no seu dia-a-dia, descrevendo seus sentimentos e emoções em forma de texto. Desse modo, as palavras utilizadas podem refletir os estados psicológicos e emocionais dos indivíduos que as escrevem [Pennebaker 2013].

O conjunto de dados `MQD1016` foi criado em 2016 com textos de usuários do MQD. Ele é composto por textos de 510 usuários do sexo feminino e 506 usuários do sexo masculino. As palavras presentes nesse conjunto de dados foram preprocessadas de forma que apenas as palavras existentes na língua portuguesa fossem utilizadas. Assim, foram removidas palavras escritas de forma incorreta, como por exemplo as palavras com repetições de letras (e.g. felizzzz, Oiiii). Esse conjunto de dados contém 7480 atributos (palavras diferentes).

3.3. Extração das características dos usuários

Neste trabalho utilizamos o dicionário de palavras do LIWC em português [Filho et al. 2013] para filtrar as palavras não existentes na língua portuguesa. Com isso, o conjunto de dados `MQD1016` possui, para cada usuário, o conjunto de palavras utilizadas em seus textos que existiam no dicionario do LIWC. Vale ressaltar que todas as palavras que continham letras maiúsculas foram substituídas pela mesma palavra em caixa-baixa.

Nesse conjunto de dados, cada usuário é representado como um vetor de p posições, em que p é o tamanho do conjunto de palavras formado pelos textos de todos os usuários. Cada posição do vetor representante de um usuário contém o número de vezes que a palavras no índice x_i ocorreu nos textos desse usuário. Assim, todos os usuários possuem um vetor de tamanho p.

Primeiramente, utilizamos o `MQD1016` para a criação de um novo conjunto de dados denominado `MQD1016-LIWC-PT`, no qual cada usuário é representado por um

vetor de 64 posições. As posições do vetor de cada usuário representam a contagem de palavras (em seus textos) relativas a cada categoria do dicionário de dados do LIWC. Para produzir o vetor de um usuário u, as palavras utilizadas nos textos do respectivo usuário é pesquisada no dicionário do LIWC em português. Isso retorna, para cada palavra, um conjunto de n categorias $1 < n < m$ em que n é o número de categorias de cada palavra e m é o número total de categorias, no caso do LIWC, 64. Em seguida, as posições (no vetor) de cada categoria retornada (x_j) são incrementadas.

A Figura 1 ilustra um vetor de palavras representando um usuário do conjunto de dados `MQD1016-LIWC-PT`. Cada posição do vetor reflete o número de palavras em cada uma das 64 categorias do LIWC (x_i). Por exemplo, pode-se notar que 10 termos se enquadraram na categoria x_1. Todos os usuários são representados segundo esse modelo.

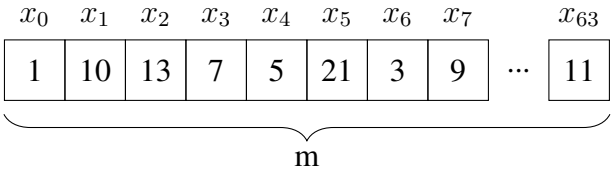

Figura 1. Vetor representando um usuário utilizando as categorias do LIWC.

Em seguida, foi gerado um terceiro conjunto de dados, denominado `MQD1016-LIWC-PT-20p`. Para gerar esse conjunto de dados, foram selecionadas 20% das 64 categorias seguindo o teorema de Pareto [Koch 1999]. Nesse cenário, foram selecionadas as 13 categorias que apresentavam maior ganho de informação (*information gain*). As categorias selecionadas foram: pronoum, ppron, i, preps, humans, negemo, ingest, relativ, space, time, leisure, money, relig, class.

Os três conjuntos de dados produzidos (i.e. `MQD2016`, `MQD2016-LIWC-PT` e `MQD2016-LIWC-PT-20p`) foram empregados para produzir modelos de classificação. Em seguida, esses modelos foram utilizados para inferência do sexo. Os experimentos foram executados utilizando a técnica de validação cruzada denominada *k-fold validation* com dez partições. A media *F1-score* foi utilizada para a avaliação dos resultados obtidos. Esses resultados são apresentados na Seção 4.

4. Resultados experimentais

Essa seção apresenta os resultados experimentais encontrados na inferência do sexo de usuários de redes sociais utilizando o LIWC em português do Brasil. Para geração desses resultados, utilizamos algoritmos de classificação conhecidos na literatura: ZeroR, Random Forest (RF), Naive Bayes (NB), NB Multinomial e SMO.

A Tabela 1 descreve os resultados obtidos. Os valores em negrito indicam o conjunto de dados que obteve o melhor F1 para os algoritmos apresentados. O algoritmo ZeroR foi utilizado como *baseline*. Pode-se notar que os algoritmos Random Forest RF, NB e NB Multinomial apresentaram melhor F1 para o conjunto de dados `MQD1016`. O algoritmo SMO apresenta melhores resultados para o conjunto de dados `MQD1016-LIWC-PT-20p`.

É importante ressaltar que, embora os algoritmos RF, NB e NB Multinomial apresentem F1 inferiores nos conjuntos de dados `MQD1016-LIWC-PT` e

Tabela 1. Classificação de sexo - Média F1

	ZeroR	*RF*	*NB*	*NB Multinomial*	SMO
MQD1016	0.336	**0.645**	**0.611**	**0.656**	0.602
MQD1016-LIWC-PT	0.336	0.574	0.565	0.566	0.598
MQD1016-LIWC-PT-20p	0.336	0.599	0.582	0.565	**0.607**

`MQD1016-LIWC-PT-20p`, muitas palavras do conjunto de dados `MQD1016` são flexionadas em gênero, o que pode influenciar o resultado da classificação, Essa característica pode ser observada ao selecionarmos as palavras que apresentam maior ganho de informação. As 10 palavras que produziram maior ganho de informação no conjunto de dados `MQD1016` são: ele, cansado, ela, para, sozinha, dele, eu, meu, mais, menino. Pode-se notar três adjetivos flexionados em gênero (i.e. cansado, sozinha, menino), dois pronomes possessivos que possuem traços de marcação de gênero (i.e. dele, meu) e dois pronomes pessoais que também apresentam traços de marcação de gênero (i.e. ele, ela). Apenas três palavras não possuem informações de gênero (i.e. para, eu, mais).

Nesse cenário, fica evidente que a tarefa de classificação nos conjuntos de dados produzidos com o LIWC é capaz de apresentar resultados satisfatórios, sem que haja influência direta das palavras que possuem traços de gênero. Assim, o procedimento de extração de características aqui proposto é relevante, visto que, um dos algoritmos utilizados (SMO) apresentou melhores resultados para o conjunto de dados `MQD1016-LIWC-PT-20p`. Essa evidência fica mais clara quando removemos as 7 palavras com traços de gênero (i.e. cansado, sozinha, menino, dele, meu, ele, ela) do conjunto de dados `MQD1016`, o que totaliza uma remoção de menos de 0.01% dos 7480 atributos (palavras). Ao executar novamente o algoritmo que apresentou o melhor resultado para esse conjunto de dados (i.e. NB Multinomial), houve uma perda de 3% na média F1.

5. Conclusões e trabalhos futuros

Nesse trabalho, apresentamos uma avaliação da inferência do sexo de usuários de redes sociais utilizando o dicionário de dados do LIWC para o português do Brasil. Essa avaliação tem o objetivo de contribuir na lacuna de trabalhos que envolvem o dicionário mencionado, visto que sua versão inicial foi proposta recentemente em [Filho et al. 2013].

Para isso, produzimos três conjuntos de dados utilizando textos provenientes de uma rede social brasileira. Avaliamos esses três conjuntos de dados com quatro algoritmos: RF, NB, NB Multinomial e SMO. De forma geral, os resultados obtidos foram satisfatórios. Vale ressaltar que para o caso do algoritmo de classificação SMO, os melhores resultados foram alcançados com o conjunto de dados `MQD1016-LIWC-PT-20p`. Esses resultados podem servir como base para trabalhos que realizem classificação de textos em português do Brasil. Os resultados preliminares alcançados indicam bons resultados.

É interessante ressaltar que alguns trabalhos que realizam a inferência de sexo em textos são sensíveis a palavras flexionadas em gênero. Assim, palavras como `cansada`, `obrigado`, `obrigada`, poderiam ter forte influência na inferência do sexo. Essas palavras são evidenciadas em [Araújo et al. 2014] como parte de um grupo de 20 palavras (atributos) que possuem maior ganho de informação na separação entre as duas classes do sexo dos usuários. Por outro lado, não existe no LIWC uma classe sensível ao gênero. As-

sim, o que define os resultados apresentados no decorrer desse trabalho são características de preocupações pessoais, processos linguísticos, psicológicos, etc.

Durante o desenvolvimento desse trabalho, emergiram algumas ideias para trabalhos futuros, dentre elas, a inferência da idade de usuários utilizando o dicionário do LIWC em português do Brasil. Além disso, seria interessante que alguns esforços fossem investidos para o estudo da inferência da personalidade de usuários utilizando esse mesmo dicionário. Conforme mencionado, já existem trabalhos nesse âmbito utilizando o dicionário do LIWC em inglês. Vale ressaltar que, como trabalho futuro, iremos realizar a comparação com outras abordagens da literatura.

Referências

[Araújo et al. 2014] Araújo, M., Gonçalves, P., and Benevuto, F. (2014). Métodos para análise de sentimentos no twitter. In *CSBC 2014 - BraSNAM ()*.

[Filho et al. 2013] Filho, P. P. B., Pardo, T. A. S., and Aluísio, R. M. (2013). An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis.

[Filho et al. 2014] Filho, R. M., Carvalho, A. I. R., and Pappa, G. L. (2014). Inferência de sexo e idade de usuários no twitter. In *CSBC 2014 - BraSNAM ()*.

[Golbeck et al. 2011] Golbeck, J., Robles, C., Edmondson, M., and Turner, K. (2011). Predicting personality from twitter. In *SocialCom/PASSAT*, pages 149–156. IEEE.

[Goswami et al. 2009] Goswami, S., Sarkar, S., and Rustagi, M. (2009). Stylometric analysis of bloggers' age and gender. In *Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009, San Jose, California, USA, May 17-20, 2009*.

[Koch 1999] Koch, R. (1999). *The 80/20 Principle: The Secret of Achieving More with Less*. A Currency book. Doubleday.

[Kokkos and Tzouramanis 2014] Kokkos, A. and Tzouramanis, T. (2014). A robust gender inference model for online social networks and its application to linkedin and twitter. *First Monday*, 19(9).

[Nguyen et al. 2013] Nguyen, D., Gravel, R., Trieschnigg, D., and Meder, T. (2013). "how old do you think i am?": A study of language and age in twitter. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, ICWSM 2013.

[Peersman et al. 2011] Peersman, C., Daelemans, W., and Van Vaerenbergh, L. (2011). Predicting age and gender in online social networks. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, SMUC '11, pages 37–44, New York, NY, USA. ACM.

[Pennebaker 2013] Pennebaker, J. (2013). *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury USA.

[Pennebaker et al. 2001] Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). *Linguistic Inquiry and Word Count*. Lawerence Erlbaum Associates, Mahwah, NJ.

[Schler et al. 2006] Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. (2006). Effects of Age and Gender on Blogging. In *Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*.

[Schwartz et al. 2013] Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., and Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, 8(9):e73791.