Inferência de Sexo e Idade de Usuários no Twitter *

Renato Miranda Filho[1,2], **Arthur I. R. Carvalho**[1], **Gisele L. Pappa**[1]

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)
CEP 31270-010 – Belo Horizonte – MG – Brasil

[2]Instituto Federal de Minas Gerais (IFMG)
CEP 34515-640 – Sabará – MG – Brasil

{renato.miranda,iperoyg,glpappa}@dcc.ufmg.br

Abstract. *Online social networks have attracted a large audience of users, who take advantage of their tools to discuss all kinds of subjects. Several studies have been conducted to identify the topics discussed, and there is a growing number of studies focusing on the personal characteristics of the users participating in these discussions. This study is interested in the gender and age of users, and as many others, relies on the messages posted by the users to infer their characteristics. Previous studies have been proposed to infer the gender and age of the individuals, but mostly in English. We develop an approach that deals with the Portuguese language. The methods developed for gender and age inference obtained an approximate accuracy of 90% and 80%, respectively.*

Resumo. *As redes sociais online têm atraído um grande público de usuários, que utilizam suas ferramentas para a discussão dos mais variados assuntos. Diversos trabalhos já foram realizados para identificar os temas discutidos nas redes, e um número crescente de trabalhos focando nas características pessoais dos usuários que participam dessas discussões vem sendo desenvolvidos. Este trabalho se propõe a inferir o sexo e idade de usuários da rede Twitter. Alguns trabalhos já se propuseram a fazer o mesmo, mas não foram encontrados estudos que se concentram no público que utiliza a língua portuguesa em suas mensagens. Os métodos para inferência do sexo e idade desenvolvidos neste trabalho alcançaram acurácias aproximadas de 90% e 80%, respectivamente.*

1. Introdução

As redes sociais online têm atraído um grande público de usuários, que as utilizam para discussão dos mais variados assuntos. Esse fenômeno vem despertando o interesse da comunidade científica que, aproveitando o grande volume de dados criado diariamente, propõe novos métodos para identificar eventos [Yang and Leskovec 2011, Cataldi et al. 2010, Lin et al. 2011], sentimentos [Turney 2002, Hu et al. 2013, Gonçalves et al. 2013] e opiniões [Asur and Huberman 2010, Tumasjan et al. 2010] expressas nessas redes.

Tão importante quanto identificar o assunto que permeia tais mensagens é conseguir individualizar os usuários, ou seja, extrair suas características pessoais sem se beneficiar de dados privados. Informações deste tipo podem ser úteis, por exemplo, para uso em

*Este trabalho foi parcialmente financiado pela Fapemig, CNPq, CAPES e InWeb.

campanhas de publicidade direcionadas a públicos específicos ou em estudos sociológicos sobre determinadas camadas da população.

Algumas redes impõem para o cadastro do usuário o preenchimento de informações que indiquem, por exemplo, o sexo, a idade, a localização geográfica, entre outros. Nem sempre estes conteúdos estão disponíveis ou são públicos. Por exemplo, em [Gundecha et al. 2011] os autores mostraram que em uma amostra de aproximadamente 2 milhões de usuários do Facebook 82% informaram o sexo e somente 30% informaram sua cidade atual, tornando propício o desenvolvimento de mecanismos capazes de inferir tais informações.

Diversos trabalhos já se propuseram a identificar o sexo e idade do usuário, sendo o segundo considerado uma tarefa mais complexa. No entanto, não foram encontrados trabalhos que dirigissem sua abordagem a identificação de tais características para usuários que postam conteúdos em português. A linguagem utilizada é de fundamental importância nos métodos desenvolvidos, pois estes são normalmente baseados no texto publicado pelos indivíduos e, no caso da inferência de sexo, no nome do usuário. Este trabalho vem preencher tal lacuna.

Assim, foram desenvolvidos métodos para inferência do sexo e idade, que alcançaram acurácias aproximadas de 90% e 80%, respectivamente. Para sexo a técnica desenvolvida é constituída de duas etapas executadas sequencialmente: verificação do nome do usuário em um dicionário com rótulos previamente definidos e um algoritmo supervisionado de classificação, treinado com mensagens de usuários que utilizaram um conjunto de termos identificados como relevantes para o contexto. Já para idade a estratégia leva em consideração o conjunto das últimas mensagens postadas pelos usuários coletados, de cada faixa etária de interesse, como treinamento para um algoritmo de classificação supervisionada.

O artigo está estruturado da seguinte forma, a Seção 2 apresenta alguns trabalhos que serviram de base para este artigo. A Seção 3 descreve a metodologia empregada no trabalho. A Seção 4 trata especificamente da inferência do sexo e a Seção 5 do método para inferência da idade. Por fim, a Seção 6 apresenta as conclusões e propostas de continuidade deste trabalho.

2. Trabalhos Relacionados

A maioria dos trabalhos que focam na inferência de sexo e idade utilizam as mensagens postadas para extrair características textuais típicas de determinados gêneros ou faixas etárias. Estas características podem representar, por exemplo, um maior uso de termos pertencentes a uma determinada classe gramatical ou mesmo aspectos específicos do ambiente virtual como, por exemplo, a quantidade de *hiperlinks* publicados.

Partindo de uma análise de mensagens postadas em milhares de blogs, o trabalho realizado por [Schler et al. 2006] identificou algumas diferenças nos conteúdos coletados como, por exemplo, homens escrevem mais sobre política, tecnologia e dinheiro e mulheres mais sobre vida pessoal. No estilo de escrita os autores identificaram, por exemplo, que mulheres usam mais pronomes, palavras de negação e concordância enquanto homens usam mais artigos e preposições. Outras características, como o uso de *hiperlinks* e número de palavras, também se mostraram como fator de diferenciação entre sexos,

onde o primeiro ocorre mais entre os homens e o segundo mais entre as mulheres. Estas diferenças também se refletem nas faixas etárias em que, de forma geral, observa-se que as mensagens tendem a adquirir um comportamento mais masculino entre os grupos mais velhos.

Ainda no universo dos blogs, [Goswami et al. 2009] exploraram características textuais do tipo uso de gírias e a variação no comprimento médio das sentenças. Com isso, observou-se que um maior uso de gírias e mensagens mais curtas são características encontradas em conteúdos postados por usuários mais jovens. Após montar um conjunto de treinamento foi realizada uma classificação pelo método *Naive Bayes*, obtendo uma precisão de aproximadamente 80% na identificação de gênero e de aproximadamente 90% na identificação da faixa etária.

Evoluindo o âmbito de pesquisa para os microblogs, um dos primeiro trabalhos foi realizado por [Rao et al. 2010]. Para obter as características pessoais dos usuários do Twitter os autores coletaram informações disponibilizadas pelos indivíduos em outros meios como, por exemplo, blogs. Alguns dos desafios citados pelos autores para estudos neste ambiente são o tamanho limitado dos textos, a natureza informal da linguagem utilizada, a falta de sinais prosódicos e a ausência de sinais sociolinguísticos tradicionais.

[Peersman et al. 2011] realizaram um estudo exploratório, com uma base coletada da rede social belga Netlog. No quesito idade, o principal objetivo dos autores era classificar adultos e adolescentes e, com isso, auxiliar na tarefa de identificar possíveis usuários pedófilos. Já o quesito sexo é de interesse dos autores, porque existe uma predominância em pessoas do sexo masculino entre os indivíduos pedófilos. Assim, foi utilizado um classificador SVM e a técnica para seleção de atributos χ^2(*Chi Square*). Foi observado que a escolha de palavras, tais como "bro"(*brother*) e "grts" (*greetings*) parece ser mais importante para a previsão da idade do que a forma como tais termos são combinados. O SVM conseguiu obter resultados promissores na identificação de adolescentes e adultos, principalmente em faixa etária maiores. A informação de gênero também se mostrou útil na construção de um classificador de idade mais preciso.

Outro estudo sobre a relação existente entre o uso da linguagem e a previsão da idade dos usuários do Twitter, utilizando contas holandesas, é realizado em [Nguyen et al. 2013]. O principal fator abordado são as mudanças que ocorrem com a diferença etária. Neste trabalho, os autores trabalham com a classificação de idade em três níveis: (i) faixas etárias; (ii) fases da vida; e (iii) idade exata. Algumas das observações realizadas foram: a relevância de identificar o sexo dos indivíduos em trabalhos para previsão de idade, o comportamento de pessoas mais jovens, com alongamento de palavras (repetição de caracteres) e uso constante da primeira pessoa, e o de pessoas mais velhas, com *tweets* mais longos e maior uso de preposições.

3. Metodologia

Os métodos analisados para os processo de caracterização foram baseados nos textos publicados pelos usuários e em atributos não textuais extraídos dos *tweets*, estratégia similar a adotada por diversos trabalhos encontrados na literatura [Nguyen et al. 2013, Mahmud et al. 2012, Cheng et al. 2010]. A metodologia foi dividida em 5 etapas:

1. Coleta de *tweets*: foram coletados, utilizando a API do Twitter, os 200 *tweets* mais recentemente publicados pelos usuários identificados. Usamos os últimos

200, pois é um limite que a API do Twitter nos permite coletar de forma mais simplificada. De forma geral, esses usuários possuem duas características principais: publicaram *tweets* públicos e em Português.
Para identificarmos os usuários coletados utilizamos duas abordagens distintas: (i) usuários que postaram mensagens relacionadas com as Eleições 2012 ou com o programa BBB 13 e (ii) usuários que utilizaram um termo comum no idioma Português, de forma semelhante à abordagem seguida em [Nguyen et al. 2013].
O termo escolhido para orientar o processo de rastreamento foi *coisa*. De acordo com um documento publicado pela Academia Brasileira de Letras[1], coisa é o substantivo mais frequente usado no Português Brasileiro. Observe que primeiro coletamos os *tweets* e deles encontramos os usuários para, então, recuperar seus últimos 200 *tweets*.

2. Pré-processamento das mensagens: foram retiradas acentuações gráficas, pontuações, termos *stop words*, caracteres não ASCII e considerados todos os caracteres em minúsculo.
3. Verificação da importância dos termos utilizados: avaliamos os termos encontrados utilizando as métricas de Ganho de Informação ou pelo *Chi Squared*(χ^2).
4. Classificação de usuários utilizando algoritmos supervisionados: resultados com quatro classificadores diferentes são apresentados: *Naive Bayes Multinomial*, *Naive Bayes*, SVM e *Random Forest*. Todos os resultados foram obtidos usando as versões Weka destes classificadores [Witten and Frank 2005]. O *Naive Bayes Multinomial* (NBM) foi escolhido por ser extremamente rápido e apresentar bons resultados com texto, e o SVM e *Random Forest* por estarem entre os classificadores estado da arte. Em todos os classificadores, exceto o SVM, foram utilizados os parâmetros padrão. Os parâmetros para o SVM foram otimizados usando o ferramenta *easy*, que realiza uma pesquisa de grade na escolha dos valores. Os experimentos foram realizados utilizando um procedimento de validação cruzada de cinco partições.
5. Avaliação dos resultados: os resultados serão avaliados conforme valores encontrados para acurácia e média F1, métrica apropriada para avaliação em conjuntos desbalanceados.

Especificidades sobre cada caracterização serão discutidas nas seções a seguir.

4. Identificação de sexo

Queremos identificar o sexo do usuário, ou seja, feminino ou masculino. Para tanto, o método proposto se baseia em duas etapas: dicionário de nomes e classificação por algoritmos supervisionados.

4.1. Fase 1: Dicionário de nomes

Para construir o dicionário de nomes, utilizamos uma base coletada na rede social Facebook[2] constituída por nomes de usuários e seus respectivos sexos. O Facebook foi utilizada porque tem uma maior disponibilidade de nome e sexo de usuários. Essa base foi enriquecida utilizando o catálogo de nomes para bebês contido no sítio BebeAtual[3].

[1] http://www.academia.org.br/
[2] http://www.facebook.com/
[3] http://bebeatual.com/

Cada nome foi identificado com as informações do sexo (masculino, feminino ou unissex) e a frequência em que o nome foi encontrado para cada um deles.

O dicionário gerado foi então pré-processado, descartados duplicatas, nomes com menos de três caracteres e desconsiderados a informação de sobrenomes. Como resultado final foi criado um dicionário com 21,378 nomes femininos, masculinos e unissex. Mantivemos na base nomes unissex quando a frequência em algum dos sexos fosse dez vezes superior a do outro, gerando um total de 20,801 nomes, sendo 11,671 femininos e 9,130 masculinos.

Considerando que o dicionário criado fornecia uma base pela qual poderíamos classificar os usuários com alta confiança e média cobertura, criamos 3 bases de dados rotuladas para serem utilizadas na fase de classificação: (i) base genérica: usuários que postaram mensagens no contexto das eleições 2012 ou BBB 13; (ii) seguidores do perfil *Mens Health* Brasil[4]; e (iii) seguidores do perfil *Womens Health* Brasil[5]. As duas últimas fontes de usuários escolhidas por apresentarem contextos tipicamente masculinos e femininos, respectivamente.

Dentre os conjuntos de usuários coletados foram selecionados somente aqueles em que o nome foi identificado pelo dicionário de nomes como sendo do sexo masculino ou feminino, para os quais foi possível coletar no mínimo 50 *tweets*. Tais *tweets* foram pré-processados. Após isso, para cada usuário foi formado um vetor para identificar a presença das palavras mencionadas, com frequência mínima de três.

4.2. Fase 2: Classificação Supervisionada

As 3 bases criadas anteriormente foram utilizadas para gerar modelos de classificação, e os resultados obtidos são mostrados na Tabela 1. Podemos observar pelos F1 encontrados que o SVM e o *Random Forest* obtiveram resultados muito inferiores em classes altamente desbalanceadas, como é o caso das bases *Mens Health* e *Womens Health*. Já o *Naive Bayes* possui resultados inferiores ao *Naive Bayes* Multinomial. Assim, o algoritmo *Naive Bayes* Multinomial será utilizado daqui em diante.

Tabela 1. Classificação de sexo - Média F1 utilizando todos os termos (H= masculino e F=feminino)

Base	Tamanho da base	SVM	*Random Forest*	*Naive Bayes* (NB)	NB Multinomial
Womens Health (W)	836 (H) e 2,971 (F) = 3,807	0.438	0.520	0.622	0.655
Mens Health (M)	9,513 (H) e 2,180 (F) = 11,693	0.450	0.549	0.598	0.637
Revistas (W+M)	10,349 (H) e 5,151 (F) = 15,500	0.748	0.669	0.711	0.727
BBB 13	518 (H) e 788 (F) = 1,306	0.617	0.578	0.682	0.681
Eleições	3,759 (H) e 2,535 (F) = 6,294	0.774	0.658	0.682	0.690
Completa	14,626 (H) e 8,474 (F) = 23,100	0.775	0.678	0.701	0.703

Para tentarmos melhorar os resultados obtidos, utilizamos o Ganho de Informação e χ^2 para verificar o melhor número de atributos (palavras) capazes de separar as duas classes de sexo dos usuários, para quantidades variando de 10 em 10. Para tanto, utilizamos a base completa de usuários coletados, 23,100 indivíduos.

Como mostrado na Tabela 2, com 20 termos somos capazes de obter boa média F1 e, ao mesmo tempo, uma alta, já que 97.25% dos usuários utilizou no mínimo um dos

[4]https://twitter.com/menshealth_br

[5]https://twitter.com/WomensHealthBR

termos selecionados. As termos são listados abaixo:

- obrigada, obrigado, cansada, lindo, amei, adorei, adoro, amiga, amo, lt (abreviação para *Last tweet*), gol, futebol, delicia, linda, @hugogloss, lindos, libertadores, campeao, jogador e palmeiras.

Como podemos observar, os termos selecionados, em geral, apresentam sufixos masculinos (“o”) ou femininos (“a”). Além disso, termos tipicamente atribuídos a universos distintos de gêneros também foram selecionados, como palavras relacionados ao futebol.

Tabela 2. Sexo Base Completa - seleção de atributos

# atributos	Usuários que utilizaram os termos (%)	F1		
		Homem (14,626)	Mulher (8,474)	Média
10	93.39	0.836	0.593	0.714
20	97.25	0.812	0.651	0.732
30	98.83	0.796	0.655	0.725
40	99.36	0.785	0.647	0.716
50	99.77	0.789	0.655	0.722
60	99.77	0.781	0.650	0.715
70	99.82	0.781	0.653	0.717
80	99.88	0.780	0.654	0.717
90	99.88	0.779	0.655	0.717
100	99.88	0.774	0.655	0.714

4.3. Avaliação do método

O método proposto consiste na execução em sequência de duas fases: verificação do nome no dicionário e, caso o nome do usuário não esteja rotulado, classificação por um algoritmo supervisionado.

Inicialmente, para avaliar a eficácia do dicionário isolamos, dentre os usuários coletados na base das Eleições e do BBB, aqueles que indicavam *links* para *blogs* na rede social Blogspot[6], com informação do sexo do usuário. Verificamos então a compatibilidade do sexo informado pelos usuários coletados com o que foi catalogado. para esse subconjunto de usuários, o dicionário acertou 98% dos nomes analisados.

Considerando então a base completa, avaliamos: (i) o dicionário; (ii) o algoritmo supervisionado; e (iii) p método completo (dicionário + algoritmo supervisionado). Quando avaliamos o dicionário separadamente, atribuímos o sexo “masculino” para todos os nomes não catalogados. Os resultados obtidos são mostrados na Tabela 3, em que na base coletada dos blogs BBB foram analisados 1,293 usuários e na base das eleições foram analisados 5,875 usuários. Assim, consideramos que o método completo, ou seja, dicionário seguido pela avaliação pelo *Naive Bayes* Multinomial (20 atributos) é o mais apropriado para a classificação do sexo.

Tabela 3. Acerto das variações do método (%)

	BBB13	Eleições
Dicionário	74.40	87.76
Naive Bayes Multinomial (20 atributos)	78.11	80.14
Método completo	90.02	92.54

[6] `http://www.blogger.com/`

5. Identificação de idade

Diversos trabalhos já abordaram a identificação de idade e verificaram que tanto a forma de linguagem dos indivíduos, como a maior recorrência no uso de preposições ou artigo, quanto as relações interpessoais, como a criação de vínculos de amizades, são características capazes de diferenciar indivíduos de diferentes faixas etárias.

Assim, inicialmente mostraremos a estratégia utilizada para formação da nossa base de dados, identificação e avaliação do poder discriminativo de atributos textuais e não textuais e, por fim, um estudo sobre o poder preditivo de alguns classificadores estado da arte nesta tarefa.

5.1. Base de dados

A base de dados utilizada nesta etapa foi construída a partir de dois processos distintos, são eles:

1. Coleta em blogs: Usuários identificados nas coletas BBB 13, Eleições 2012 ou que usaram a palavra "coisa" em suas mensagens, cujos *tweets* continham indicação para um blog da rede Blogspot.
   Assim, foi realizado um *crawler* sobre tais blogs e extraída as idades dos usuários pelo campo descrição com a expressão "X anos", onde X representa um valor entre 1 e 99. Para garantir que a expressão continha informação da idade do indivíduo, cada texto foi analisado e filtrado manualmente.
2. Coleta manual: Foi utilizada a ferramenta de busca do Twitter[7] e por meio de pesquisas pelos termos "tenho X anos" ou "fiz X anos", onde X representa um valor entre 10 e 99, analisou-se manualmente se a mensagem tratava de uma experiência pessoal e se a fotografia apresentada no perfil era condizente com a idade mencionada, identificando, desta forma, usuários com suas respectivas idades.
   Além disso, para a classe de usuários com idade superior a 45 anos, foram realizadas pesquisas pelos termos "sou aposentado", "aposentado", "aposentadoria", "netinhos" ou "cobap" (Confederação dos Aposentados e Pensionistas do Brasil) e por seguidores dos usuários de *screen names* "terceira_idade", "blogda3idadesp", "aTerceiraIdade" e "nucleo3idade". Tais indivíduos foram adicionados à base quando as fotografias apresentadas no perfil evidenciavam uma pessoa com aparência condizente a esta faixa etária.

Como resultado, obteve-se uma base com 1,709 usuários rotulados. Finalmente, foram coletados os 200 últimos *tweets* de cada usuário identificado.

5.2. Extração de características

Tendo uma base de dados com usuários rotulados, podemos identificar as características que são capazes de distinguir pessoas de diferentes faixas etárias. Analisaremos basicamente dois grupos de características: textuais e não textuais.

5.2.1. Atributos textuais

Para extrair características textuais, utilizamos os 200 *tweets* mais recentes postados pelos usuários. Inicialmente, consideramos os atributos mais discriminativos em todo o

[7] https://twitter.com/search-home

texto. Em uma segunda fase, com base em nossa intuição, observação das palavras mais frequentemente utilizadas em cada idade e seguindo exemplos encontrados em trabalhos relacionados, identificamos um conjunto de termos característicos de cada faixa etária, retratando eventos, lugares e expressões típicas de cada etapa da vida.

Nesta segunda fase, para cada característica considerada relevante um conjunto de termos foi manualmente definido e suas frequências médias e medianas foram avaliadas. Foram criados conjuntos de termos nos seguintes âmbitos: (i) Filmes adolescentes; (ii) Ídolos adolescentes; (iii) Programas de televisão para adolescentes; (iv) Bar; (v) Bebidas; (vi) Faculdade; (vii) Maquiagem; (viii) Abreviações; (ix) Diversão noturna; (x) Política; (xi) Relacionamento; (xii) Religião; (xiii) Escola; (xiv) Casamento; e (xv) Termos frequentes utilizados por pessoas com mais de 45 anos. Uma lista completa desses termos está disponível em [8].

Além destes termos, para cada faixa etária considerada, definimos um conjunto de gírias e internetês normalmente utilizadas. Como dicionário inicial destas palavras coletamos termos do blog Dicionariopopular[9], um catálogo de gírias e expressões populares, e de Linguadedoido[10], um catálogo de internetês. Tais catálogos também passaram pelo pré-processamento.

Para atribuirmos cada palavra dos dicionários às faixas etárias, seja ela gíria ou internetês, calculamos o tf-idf (*term frequency-inverse document frequency*) de cada termo nos documentos constituídos pela concatenação dos textos de todos usuários de cada faixa definida. Foram considerados os 20 termos com maior valor para esta métrica, que se diferenciavam dos 20 melhores termos das demais faixas etárias. Uma lista completa desses termos está disponível em [11].

Outros atributos textuais analisados foram:

- Média e mediana de termos positivos: conjunto de palavras manualmente selecionadas e expandidas pelo dicionário de sinônimos Dicio[12]. São termos como: bom, admiro e adoro.
- Média e mediana de termos negativos: conjunto de palavras criado de forma análoga às positivas. São termos como: péssimo, raiva e desprezível.
- Média e mediana de erros ortográficos;
- Média e mediana do uso de: preposições, artigos e *emoticons*;

Para exemplificar o poder discriminativo dos conjuntos de termos criados são mostrados nos gráficos das Figuras 1 e 2, as médias por usuário de menções a termos dos conjuntos "diversão noturna" e "termos frequentes mais de 45 anos", respectivamente. Estes gráficos foram construídos de forma a retratar por cada circunferência a média de menções que o usuário utilizou em termos dos conjuntos em seus *tweets* mais recentes. Foram desconsiderados usuários que não mencionaram os termos dos conjuntos, ou seja, média igual a 0. Desta forma, quanto mais escuro o ponto mais usuários possuem o mesmo comportamento.

[8] http://www.dcc.ufmg.br/~renato.miranda/brasnam_sexo_idade.html

[9] http://dicionariopopular.blogspot.com.br/

[10] http://linguadedoido.blogspot.com.br/2008/07/dicionrio-de-internets.html

[11] http://www.dcc.ufmg.br/~renato.miranda/brasnam_sexo_idade.html

[12] http://www.dicio.com.br

Figura 1. Média de menção a palavras de diversão noturna por usuário

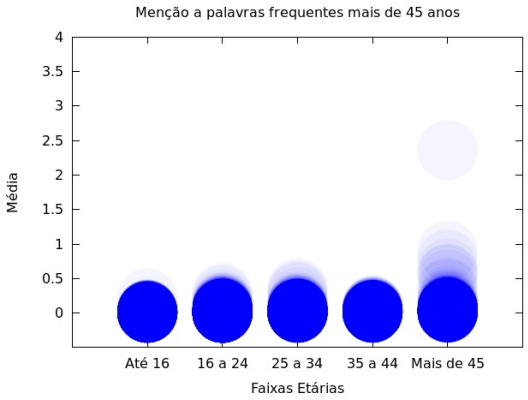

Figura 2. Média de menção a palavras frequentes mais de 45 anos por usuário

Como podemos observar, os conjuntos selecionados retratam o comportamento de usuários em diferentes etapas da vida, com maior ocorrência do primeiro conjunto entre os mais jovens e do último entre os mais velhos.

5.2.2. Atributos não textuais

Complementando os atributos textuais apresentados na subseção anterior, também extraímos e analisamos os seguintes atributos não textuais:

- Sexo: inferido conforme método apresentado anteriormente;
- Média e mediana de *hiperlinks*: URL's identificadas;
- Média e mediana de *hashtags*: indicadas pelo símbolo "#";
- Média e mediana do número de caracteres nas palavras e nos *tweets*;
- Média e mediana da frequência de *tweets* por: dia, semana e mês;
- Média e mediana do número de palavras por *tweet*;
- Divulgação de coordenadas geográficas nos *tweets* publicados;
- Preenchimento de cidade válida: considerando cidades brasileiras;
- Número de seguidores;
- Número de seguidos.

Seguindo o mesmo procedimento já explicado para os gráficos de atributos textuais o comportamento dos usuários em alguns dos atributos não textuais são mostrados nos gráficos das Figuras 3 e 4, para uso de hashtags e os tamanhos médios dos *tweets*, respectivamente. Podemos observar que pessoas mais velhas tendem a escrever palavras e *tweets* com mais caracteres, e que o uso de *hashtags* é maior entre os usuários mais jovens.

5.3. Classificação

Depois de analisar os recursos textuais e não textuais que podem distinguir os usuários de diferentes faixas etárias, esta seção utiliza esses recursos para criar modelos de classificação capazes de discriminar os usuários. Os experimentos foram realizados considerando os usuários divididos em 3 e 5 faixas de idade, sendo a de 3 com as idades

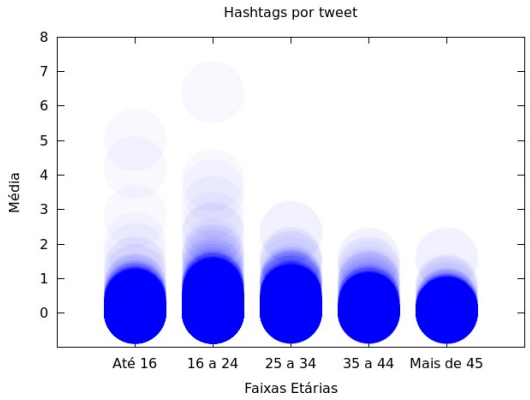

Figura 3. Média do número de *hashtags* por usuário

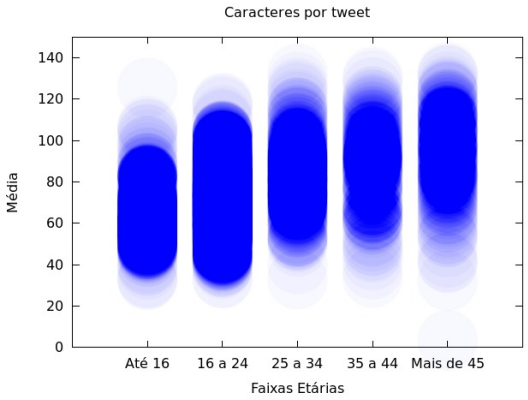

Figura 4. Média do tamanho dos *tweets* por usuário

Tabela 4. Principais termos para divisões de idade em 3 e 5 classes, conforme medida de Ganho de Informação

3 Classes	5 Classes
lt	lt (*Last tweet*)
governo	governo
uu	uu (Expressão para comemoração)
dormir	politica
politica	dormir
sdds	agr (Agora)
ne	amo
vou	sdds (Saudades)
haha	pt (Partido dos trabalhadores)
tava	mds (Meu Deus)

menor que 25, entre 25 e 45 e maior que 45 e a de 5 com as idades menor que 16, 16 a 24, 25 a 34, 35 a 44 e maior de 45.

Diferentes combinações dos recursos foram utilizadas para descrever os dados, incluindo o texto completo (em que foram observadas a presença de 5,135 termos na divisão por 3 classes e 2,859 em 5 classes, formado pelas palavras com frequência maior que 5 e Ganho de Informação maior que 0), atributos textuais selecionados, atributos não textuais selecionados e uma combinação dos atributos textuais e não textuais selecionados.

Os 10 termos mais relevantes, conforme medida de Ganho de Informação computada com o texto completo, são mostrados na Tabela 4, em que podemos observar uma grande presença de palavras relacionadas ao mundo político, abreviações e internetês.

Como podemos observar pela Tabela 5, o conjunto de classificador/atributos que conseguiu alcançar melhores resultados nos dois recortes de faixas etárias foi o *Naive Bayes Multinomial* (NBM) utilizando o texto completo postado pelos usuários. No entanto, os atributos textuais e não textuais selecionados, dado a significativa diminuição de termos avaliados, também obtiveram bons resultados quando utilizado em conjunto com o algoritmo SVM como, por exemplo, perdendo os textuais selecionados em apenas 0.1 de f1 na divisão em três classes. Isto retrata que se definidos mais cuidadosamente os termos melhores resultados poderão ser alcançados.

Tabela 5. Resultados obtidos utilizando o texto completo, atributos textuais selecionados, não textuais selecionados e todos selecionados (textuais com não textuais)

3 Classes								
	Texto completo		Textuais		Não textuais		Todos selecionados	
	$f1$	Acurácia (%)	$f1$	Acurácia (%)	$f1$	Acurácia (%)	$f1$	Acurácia (%)
NBM	0.81	81.51	0.51	64.42	0.55	64.31	0.55	64.36
Naive Bayes	0.78	77.12	0.68	66.88	0.63	67.41	0.71	71.50
SVM	0.70	74.84	0.74	75.48	0.67	70.63	0.76	76.18
Random Forest	0.66	70.51	0.71	72.79	0.66	68.46	0.71	73.26
5 Classes								
	Texto completo		Textuais		Não textuais		Todos selecionados	
	$f1$	Acurácia (%)	$f1$	Acurácia (%)	$f1$	Acurácia (%)	$f1$	Acurácia (%)
NBM	0.66	66.06	0.32	47.75	0.36	41.49	0.36	42.19
Naive Bayes	0.61	61.20	0.51	49.74	0.42	44.88	0.52	52.95
SVM	0.43	53.66	0.57	60.15	0.38	50.15	0.58	60.39
Random Forest	0.46	51.02	0.52	55.18	0.46	49.33	0.53	55.00

6. Conclusões e Trabalhos Futuros

Este trabalho abordou o problema de inferir o sexo e a idade de usuários do Twitter. Para tanto, foram utilizadas características textuais e não textuais extraídas dos conteúdos postados, estratégia similar a adotada por diversos trabalhos encontrados na literatura.

As etapas realizadas foram basicamente: identificação de indivíduos representantes de cada classe avaliada, coleta dos últimos *tweets* postados, pré-processamento do texto, verificação da importância dos termos utilizados e a classificação por diferentes algoritmos supervisionados.

Assim, foram desenvolvidos métodos para inferência do sexo e idade, que alcançaram acurácias aproximadas de 90% e 80%, respectivamente. Para sexo, a técnica desenvolvida é constituída de duas etapas executadas sequencialmente: verificação do nome do usuário em um dicionário com rótulos previamente definidos e um algoritmo supervisionado de classificação, treinado pela presença de termos identificados como relevantes no contexto encontrados nas mensagens postadas. Já para idade a estratégia leva em consideração o conjunto das últimas mensagens postadas pelos usuários.

Como trabalhos futuros pretendemos estudar métodos capazes de prover uma atualização constante e online das bases de treinamento utilizadas para as abordagens desenvolvidas. Além disso, para o contexto da identificação da faixa etária dos usuários, foi identificado um caminho promissor para classificação utilizando conjuntos de termos característicos de cada etapa da vida. Desta forma, também pretendemos identificar novos conjuntos e verificar seu impacto na classificação da idade.

Referências

Asur, S. and Huberman, B. A. (2010). Predicting the future with social media. In *Int. Conf. on Web Intelligence and Intelligent Agent Technology*, WI-IAT '10, pages 492–499.

Cataldi, M., Di Caro, L., and Schifanella, C. (2010). Emerging topic detection on twitter based on temporal and social terms evaluation. In *Int. Workshop on Multimedia Data Mining*, MDMKDD '10, pages 4:1–4:10.

Cheng, Z., Caverlee, J., and Lee, K. (2010). You are where you tweet: a content-based approach to geo-locating twitter users. In *Int. Conf. on Information and knowledge management*, CIKM '10, pages 759–768.

Gonçalves, P., Araújo, M., Benevenuto, F., and Cha, M. (2013). Comparing and combining sentiment analysis methods. In *Conf. Online Social Networks*, COSN '13, pages 27–38.

Goswami, S., Sarkar, S., and Rustagi, M. (2009). Stylometric analysis of bloggers' age and gender. In *Int. Conf. on Weblogs and Social Media*, pages 214–217.

Gundecha, P., Barbier, G., and Liu, H. (2011). Exploiting vulnerability to secure user privacy on a social networking site. In *Int. Conf. on Knowledge Discovery and Data Mining*, KDD '11, pages 511–519.

Hu, X., Tang, J., Gao, H., and Liu, H. (2013). Unsupervised sentiment analysis with emotional signals. In *Int. Conf. on World Wide Web*, WWW '13, pages 607–618.

Lin, J., Snow, R., and Morgan, W. (2011). Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In *Int. Conf. on Knowledge discovery and data mining*, KDD '11, pages 422–429.

Mahmud, J., Nichols, J., and Drews, C. (2012). Where is this tweet from? inferring home locations of twitter users. In *Int. Conf. on Weblogs and Social Media*, pages 511–514.

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 *Int. Conf. on Weblogs and Social Media*, ICWSM 2013, pages 439–448.

Peersman, C., Daelemans, W., and Van Vaerenbergh, L. (2011). Predicting age and gender in online social networks. In *Int. workshop on Search and mining user-generated contents*, SMUC '11, pages 37–44.

Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M. (2010). Classifying latent user attributes in twitter. In *Int. Workshop on Search and Mining User-generated Contents*, SMUC '10, pages 37–44.

Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. (2006). Effects of age and gender on blogging. *Symposium on Computational Approaches for Analyzing Weblogs*, pages 199–205.

Tumasjan, A., Sprenger, T., Sandner, P., and Welpe, I. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Int. Conf. on Weblogs and Social Media*, pages 178–185.

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Association for Computational Linguistics*, ACL '02, pages 417–424.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 2nd edition.

Yang, J. and Leskovec, J. (2011). Patterns of temporal variation in online media. In *Int. Conf. on Web search and data mining*, WSDM '11, pages 177–186.