Predição do resultado das eleições presidenciais do Brasil baseado em tuítes

Wilton de Paula Filho[1], Ana Cristina Bicharra Garcia[2]

[1]Instituto Federal do Triângulo Mineiro (IFTM) – Uberlândia, MG - Brazil
[2]Universidade Federal Fluminense (UFF) – Niterói, RJ – Brazil

[1]wilton.filho@iftm.edu.br, [2]bicharra@ic.uff.br

Abstract. *This work uses the context of the year 2014 Brazil's presidential election to investigate whether the winner of an election can be discovered from public messages from Twitter users. Approximately 3 million and 200 thousand messages, more than 460 million different users, with reference to the main presidential were collected and analyzed. Our results show that it is possible to estimate the outcome of elections based solely on tweets count technique. Other results also show that other techniques such as user count and sentiment analysis messages can increase the accuracy of prediction models.*

Resumo. *Este trabalho utiliza o contexto das eleições presidenciais do Brasil no ano de 2014 para investigar se o vencedor de uma eleição pode ser descoberto a partir de mensagens públicas dos usuários do Twitter. Aproximadamente 3 milhões e 200 mil mensagens, de mais de 460.000 mil usuários distintos, fazendo referência aos principais presidenciáveis foram coletadas e analisadas. Nossos resultados mostram que é possível estimar o resultado das eleições baseado apenas na técnica de contagem de tuítes. Outros resultados obtidos mostram também que outras técnicas como contagem de usuários e análise de sentimentos de mensagens podem aumentar a acurácia dos modelos de predição.*

1. Introdução

Eleições para escolha de governantes é uma parte importante na democracia de qualquer instância, seja ela um país, estado ou cidade. Através deste instrumento, cidadãos têm a oportunidade de escolher os representantes de seus desejos e necessidades. Um elemento importante numa eleição são as pesquisas de opinião pública. Lewis-Beck (2005) afirmou que o principal objetivo delas é fornecer informações aos cidadãos curiosos e também às partes interessadas para que possam fazer os ajustes apropriados caso julguem necessário.

Segundo Hillygus (2011), pesquisas de opinião pública existem desde o século XIX. Vários modelos estatísticos têm sido utilizados para prever o resultado das eleições, conforme mostrado em Lewis-Beck (2005). Os modelos propostos e utilizados, mesmo em países desenvolvidos, já apresentaram falhas na predição de resultados. Fumagalli (2011) listou vários exemplos de resultados errados produzidos por institutos de enquetes públicas *offline*, inclusive para escolha de presidentes.

Para realizar enquetes de opinião pública é necessário certo período de tempo para realização da coleta e análise dos dados, antes da divulgação dos resultados. No Brasil, por exemplo, um dos institutos de opinião pública, o Sensus, precisou de 4 dias para coletar e analisar as intenções de voto dos eleitores em relação aos presidenciáveis do segundo turno das eleições no ano de 2014 (Eleições, 2014). Além do tempo de espera para divulgação dos resultados de uma pesquisa de opinião pública, o custo para contratar um instituto para realização de enquete pública a nível nacional pode ultrapassar duzentos mil reais.

Devido ao grande volume de informações produzidas nas diversas mídias sociais, tais como Facebook, Twitter e Google+, pesquisas já têm sido conduzidas no sentido de utilizar tais informações para predição do resultado de eleições para escolha de presidentes, senadores, etc. As mídias sociais tem se tornado uma importante ferramenta popular de comunicação e interação entre as pessoas na Internet. Milhares de mensagens são postadas diariamente naquelas mídias.

O Twitter é um serviço de microbloging lançado em 2006 e tem como principal característica permitir aos seus usuários publicarem mensagens curtas de até 140 caracteres. Nestas mensagens é possível fazer menções a usuários, incorporar endereços eletrônicos de páginas web e hashtags. Em junho de 2012 os três países com o maior número de contas ativas no mundo eram os E.U.A, Japão e Brasil, respectivamente.

A primeira pesquisa sobre predição do resultado de uma eleição presidencial utilizando dados do Twitter foi iniciada por O'Connor (2010). Outros autores também utilizaram dados do Twitter para prever resultados de eleições presidenciais. Tumasjan (2010) criou um modelo para prever o resultado das eleições presidenciais na Alemanha, Choy (2011) em Singapura, Choy M. C (2012), Wong F. T. (2013), Nooralahzadeh (2013), Beauchamp (2013) e Ceron (2014) nos Estados Unidos, Nooralahzadeh (2013) e Ceron A. C. (2013) na França, Ceron A. C. (2013) e Ceron (2014) na Itália, Gaurav (2013) na Venezuela, Paraguai e Equador. Outros modelos também foram criados para prever os resultados das eleições para composição do senado americano e holandês, como mostrado em Gayo-Avello D. M. (2011) e Sang (2012), respectivamente.

Este trabalho se propõe a fazer uma revisão da literatura sobre os trabalhos já publicados a respeito do uso de mensagens públicas do Twitter para predição do resultado de eleições presidenciais no Brasil, a apresentar as etapas dos modelos mais utilizados atualmente para predição do resultado de eleições presidenciais e comparar o desempenho de cada um deles considerando o cenário político brasileiro.

2. Trabalhos relacionados e questões de pesquisa

2.1. Revisão de literatura (protocolo utilizado)

Para obter na literatura os trabalhos relacionados a esta pesquisa foram coletados dados nos portais virtuais do Google Acadêmico Brasil e Periódicos da CAPES, no recorte temporal de 2006 até maio do ano de 2015. Adotou-se como critérios de inclusão: artigos, dissertações e teses online que abordassem especificamente no título e/ou no resumo a combinação dos termos da expressão *booleana*: a OR b OR c OR d OR e OR f, onde: (a) 1 AND 5, (b) 1 AND 2 AND 5, (c) 1 AND 3 AND 5, (d) 1 AND 4 AND 5, (e) 1 AND 5 AND 6, (f) 1 AND 2 AND 5 AND 6, (g) 1 AND 3 AND 5 AND 6 e (h) 1

AND 4 AND 5 AND 6. Cada termo (1 a 6) foi definido por um conjunto de palavras-chave separadas pelo operador booleano OR: (1) Twitter OR tweets OR tweet OR Tweeting OR Social media OR rede social, (2) Presidente OR presidencial OR president OR presidential OR presidenciais OR presidential, (3) Eleição OR election OR elections OR eleições OR political OR política OR Electoral OR eleitoral, (4) Voto OR voting OR votos OR votes, (5) Brasil OR Brazil OR brazilian OR brasileiro e (6) Predição OR prediction OR predicting OR predicts OR predict OR outcome OR prever.

O exame dos dados foi realizado com análise de conteúdo temática. Uma planilha eletrônica foi utilizada para organizar os títulos e resumos dos trabalhos. Os resultados mostraram 59 produções científicas. Deste total, 47 trabalhos foram eliminados pelo título, pois não apresentavam nenhuma relação direta com o tema desta pesquisa. Em seguida, foram descartadas 5 citações e por último, foram eliminados os últimos 7 trabalhos após a análise das informações do resumo, por não apresentarem também relação direta com o tema desta pesquisa.

Utilizando as bases de dados online mencionadas, os critérios de inclusão, as palavras-chave e o protocolo proposto, pode-se concluir que ainda não há nenhuma publicação sobre o uso de mensagens públicas do Twitter para prever o resultado de eleições presidenciais do Brasil. Por este motivo, novas pesquisas bibliográficas foram realizadas com o objetivo de analisar e selecionar os principais métodos propostos e utilizados por pesquisadores para predizer os resultados de eleições presidenciais em outros países. Estes métodos serão aplicados no cenário político brasileiro e os resultados serão discutidos.

2.2. Métodos para predição

Diferentes métodos têm sido utilizados por pesquisadores para prever os resultados de eleições utilizando os dados públicos de usuários do Twitter. Estes métodos variam desde o período de coleta até o cálculo da predição. Embora alguns aspectos sejam diferentes, os métodos utilizados pelos pesquisadores para calcular a predição podem ser divididos em quatro etapas (Figura 1) coleta dos dados, filtragem dos dados, redução de viés dos dados e o cálculo da predição (Prasetyo, 2014).

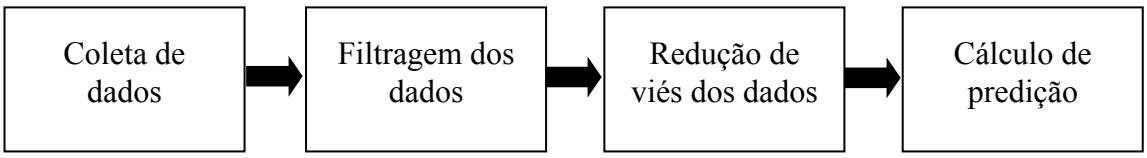

Figura 1. Modelo de predição de eleição baseado no Twitter

A coleta de dados é a etapa de obtenção de tuítes relacionados ao processo de eleição. Pesquisadores têm discutido o tempo necessário para coleta dos dados, tipos de palavras-chave a serem utilizadas para coleta dos tuítes e o método de coleta das mensagens públicas dos usuários. Wong F. M (2013), Gaurav (2013) e Mejova (2013) coletaram tuítes referentes a 6 meses antes das eleições, enquanto outros como Ceron A. C (2014) e Sanders (2013) coletaram menos de 30 dias. Períodos ainda menores foram utilizados por outros pesquisadores, como em Tumasjan (2010), Sang (2012), Cameron (2013), Nooralahzadeh (2013) e Ceron A. C (2013) que utilizaram entre 1 e 2 dias antes do dia da votação. Bermingham (2011) comparou o resultado de vários dias de dados no cálculo do modelo de predição e mostrou que 1 dia foi capaz de produzir resultados

melhores do que 5 dias de dados. Esta parece ainda ser uma questão em aberto nesta etapa do modelo do cálculo de predição. As palavras-chave utilizadas com mais freqüência para prever o resultado das eleições presidenciais tem sido o nome dos candidatos e/ou partidos. Tumasjan (2010) preveu o resultado das eleições na Alemanha utilizando esta abordagem, assim como Choy (2010) em Singapura, Gayo-Avello (2011) e Chung (2011) para prever o resultado do senado nos Estados Unidos, Gaurav (2013) na Venezuela, Paraguai e Equador. Hashtag de campanhas e de eleições também sido utilizadas como mostrado em Jungherr (2012), Nooralahzadeh (2013) e Fink (2013). Segundo Prasetyo (2014), a maioria das pesquisas tem utilizado a Search API do Twitter ao invés da Stream API, para coleta dos dados públicos dos usuários, devido a alguns fatores, como por exemplo, a alta complexidade da infraestrutura para armazenamento dos dados e ao elevado tempo necessário para realização das consultas.

Na etapa de filtragem dos dados o objetivo dos pesquisadores tem se concentrado na redução dos ruídos dos dados coletados e armazenados na base de dados. Sanders (2013) realizou a limpeza dos dados através da remoção das *stop words*, nomes dos candidatos, símbolos de pontuação, retuítes, números e endereços eletrônicos. Já Bermingham (2011) realizou a remoção de todos os tuítes que reportavam resultados das eleições. Makazhanov (2014) utilizou aprendizagem de máquina para treinar um conjunto de dados reconhecidos como spam. Chu (2010) definiu quatro testes para distinguir mensagens de humanos e robôs. Já Cook (2014) definiu nove testes para identificar tuítes falsos. Esta última abordagem foi utilizada para prever o resultado da eleição federal na Austrália em 2013. Um dos resultados obtidos por esta pesquisa foi que em apenas dois dias a quantidade de seguidores de um candidato havia aumentado mais de 40%. Testes foram realizados e os resultados comprovaram que mais de 20.000 seguidores eram robôs, mostrando assim a importância desta etapa em um modelo de predição.

Na terceira etapa do modelo pesquisadores têm desenvolvido técnicas para verificar se os usuários presentes no Twitter são uma boa representação da população. A dificuldade para chegar a esta conclusão consiste no fato dos usuários não disponibilizarem no perfil deles informações como cidade e estado onde residem, sexo, idade, escolaridade e zona (urbana ou rural) onde residem. Pesquisas já têm sido desenvolvidas com o objetivo de identificar um ou mais daqueles parâmetros. Bakker (2011) concluíram que as pessoas são frequentemente mais jovens, com nível de instrução mais alto e localizadas nos centros urbanos. Mislove (2011) compararam os usuários americanos do Twitter e a população dos Estados Unidos baseado em três informações: geografia, sexo e etnia. Uma das conclusões foi a baixíssima participação em algumas regiões, abaixo de 0,01%, enquanto outras um pouco acima de 10%. Além disso, eles perceberam que mais de 70% dos usuários eram do sexo masculino. Gayo-Avello (2011) mostrou que é possível melhorar a acurácia do modelo de predição quando informações demográficas são utilizadas. Ao considerar a idade das pessoas esta pesquisa mostrou que o *Mean Absolute Error* (MAE) ou Erro Médio Absoluto foi reduzido de 13,1% para 11,61%.

O cálculo do método de predição é a última etapa do modelo e é nela que o valor da predição é calculado. Nela são informadas a ordem dos vencedores e o percentual de votos de cada um. Segundo Prasetyo (2014), os métodos são divididos em parâmetros de contagem e análise de sentimentos. Diversos pesquisadores conseguiram encontrar

resultados satisfatórios utilizando a mera contagem de tuítes no modelo de predição, conforme encontrado em Tumasjan (2010), Ceron A. C (2013) e Sanders (2013). Sang (2012) argumentou que a contagem do número de usuários é melhor que a contagem de tuítes porque um usuário representa apenas um voto. Makazhanov (2014) já não considerou todos os tuítes na contagem, mas apenas aqueles que possuíam algum tipo de interação entre as contas dos candidatos, como por exemplo, retuítes e menções a usuários. Além da simples contagem de mensagens, autores argumentam que é importante entender o sentimento das mensagens relacionadas a cada candidato. Nesta categoria as mensagens são classificadas em positivo, negativo e/ou neutro.

Com o objetivo de verificar a possibilidade de utilização das mensagens publicadas pelos usuários do Twitter durante o período de campanha eleitoral para escolha do presidente da república do Brasil no ano de 2014, esta pesquisa irá utilizar três modelos que serão utilizados para responder cada uma das questões de pesquisa:

- O método de predição do resultado de eleições presidenciais no Brasil baseado em contagem de tuítes é comparável aos métodos de enquete pública *offline* em termos de distância do resultado real de uma eleição?
- Utilizar a contagem de usuários ao invés de tuítes melhora o resultado do modelo da predição?
- Utilizar análise de sentimento no modelo de predição é capaz de melhorar a acurácia do modelo?

3. Eleições no Brasil

A eleição para escolha do presidente do Brasil no ano de 2014 aconteceu nos dias 05 e 26 de outubro de 2014, primeiro e segundo turnos, respectivamente. Entre os dias 10 e 30 de junho de 2014 foram realizadas as convenções para a escolha dos candidatos. Onze candidatos registraram interesse em participar da corrida presidencial. O início da propaganda eleitoral gratuita no rádio e na televisão aconteceu no dia 19 de agosto. Somente após este período a população brasileira pode conhecer todos os candidatos e suas propostas. Durante a campanha eleitoral, diversas pesquisas de enquete de opinião pública *offline* foram realizadas por institutos. A Tabela 1 mostra os resultados das últimas pesquisas realizadas por quatro institutos *offline* no Brasil antes do primeiro e segundo turnos e os resultados reais da eleição para os três candidatos mais populares nas pesquisas. O MAE foi utilizado para calcular o erro de predição de cada instituto.

O instituto com o melhor desempenho no primeiro turno foi o MDA Pesquisa e no segundo turno foi o Datafolha. Considerando os dois turnos, o instituto com o melhor desempenho foi o Datafolha, com MAE médio equivalente a 1,96%.

4. Metodologia e base de dados

A primeira etapa do modelo consiste na coleta de mensagens públicas do Twitter. Para realizar esta coleta foi desenvolvida uma interface *web* utilizando a linguagem de programação PHP. O *tweet*, cidade onde o usuário reside, o número de seguidores dele e o número de menções a usuários nas mensagens, data de coleta do tweet, entre outras foram armazenados em um banco de dados. O método *search* da API do Twitter foi utilizado para coletar mensagens contendo o primeiro nome e sobrenome dos três presidenciáveis com a maior popularidade desde o início das campanhas eleitorais: Dilma Rousseff, Aécio Neves e Marina Silva. A coleta iniciou-se logo após o início da

campanha eleitoral na televisão e no rádio, isto é, no dia 21 de agosto de 2014, 45 dias antes do primeiro turno, e terminou no dia 26 de outubro do mesmo ano, dia em que foi realizado o segundo turno. Uma visão da base de dados é mostrada na Tabela 2.

No primeiro turno foram coletados mais de 3 milhões e 200 mil tuítes, uma média de 71.400 tuítes por dia. Já no segundo turno foram coletados aproximadamente 1 milhão e 500 mil tuítes, uma média de 78.557 mensagens por dia. Durante o primeiro turno 484.114 usuários diferentes publicaram pelo menos uma mensagem sobre um dos três candidatos à presidência. Já no segundo turno 376.179 pessoas publicaram pelo menos uma mensagem sobre os dois candidatos classificados para o segundo turno das eleições. Mensagens contendo nomes de dois ou mais candidatos não foram utilizadas nas análises.

Tabela 1. Resultados de enquetes públicas e o real da eleição nos dois turnos

Candidato	Datafolha		Ibope		MDA Pesquisa		Sensus[2]		Resultado real	
	1º T[1]	2º T[3]	1º T[1]	2º T[3]	1º T[1]	2º T[3]	1º T[2]	2º T[3]	1º T	2º T
Dilma R.	44%	52%	46%	53%	40,6%	49,7%	37,3%	52,1%	41,59%	51,64%
Aécio N.	26%	48%	27%	47%	24%	50,3%	20,6%	47,9%	33,55%	48,36%
Marina S.	24%	-	24%	-	21,4%	-	22,5%		21,3%	-
Preveu correto o ganhador?	Sim	Sim	Sim	Sim	Sim	Não	Não	Não	-	
MAE	4,21%	0,36%	4,54%	1,36%	3,54%	3,54%	5,77%	5,77%	-	
	1,96%		2,95%		3,54%		5,77%		-	

Na etapa do cálculo de predição do modelo foram consideradas as seguintes estratégias de contagem: mensagens e usuários. Além destas, também foi utilizada a análise de sentimentos. Foi utilizado um algoritmo de aprendizagem de máquina supervisionado para treinar e classificar as mensagens. O MAE foi a medida estatística utilizada para informar o quão próximo o resultado obtido pelo modelo de predição esteve próximo do resultado real.

Tabela 2. Alguns parâmetros da base de dados

	1º turno[4]			2º turno[5]	
	Dilma	**Aécio**	**Marina**	**Dilma**	**Aécio**
Número total de tuítes	1.281.453	540.957	1.390.685	688.151	804.449
Percentual de retuítes	59%	59,7%	54%	61,1%	63,6%
Número de usuários que publicaram mensagens sobre o candidato	236.312	100.428	242.005	179.216	170.341

5. Análises e resultados

Para responder a primeira questão de pesquisa, foram utilizados tuítes coletados um dia antes do primeiro e segundo turnos, conforme em Prasetyo (2014). Apenas os resultados

[1] Pesquisa divulgada no dia 04/10/14
[2] Pesquisa divulgada no dia 03/10/14
[3] Pesquisa divulgada no dia 25/10/14
[4] 1º turno (período de análise): de 21 de agosto de 2014 à 04 de outubro de 2015
[5] 2º turno (período de análise): de 06 de outubro de 2014 à 25 de outubro de 2015

dos três candidatos mais populares nas pesquisas *offline* foram utilizados nesta análise, por isso eles tiveram que ser normalizados. Foram considerados dois tipos de contagem de mensagens: tuítes com e sem repetição. Os resultados são apresentados na Tabela 3.

Tabela 3. Resultado do modelo de predição baseado em contagem de tuítes

Presidenciáveis/ Resultado do modelo de predição	Volume de tuítes (com repetição)		Volume de tuítes (sem repetição)		Resultado real (votos válidos)	
	1º turno	2º turno	1º turno	2º turno	1º turno	2º turno
Dilma	49,28%	52,33%	46,27%	53,96%	43,11%	51,64%
Aécio	26,44%	47,67%	26,75%	46,04%	34,78%	48,36%
Marina	24,28%	-	26,98%	-	22,11%	-
Preveu correto o ganhador?	Sim	Sim	Não	Sim	-	-
MAE	5,53%	0,695%	5,34%	2,32%	-	-

Quando tuítes repetidos foram utilizados, o modelo de predição baseado na contagem de tuítes acertou a classificação em ambos os turnos. Ao remover os tuítes repetidos o modelo não conseguiu prever corretamente os candidatos classificados para o segundo turno. Ao comparar os resultados obtidos por este modelo, sem tuítes repetidos, com os institutos *offline* nota-se que no primeiro turno ele foi superior apenas ao instituto Sensus, porém este último preveu incorretamente os classificados para o segundo turno. Já no segundo turno o modelo foi superior a quatro institutos *offline* e inferior apenas ao Datafolha, melhor colocado entre os institutos *offline*. O MAE obtido pelo modelo no segundo turno (0,695%) foi praticamente equivalente a metade do MAE do terceiro colocado (Tabela 4). Estes resultados se assemelharam ao de pesquisas anteriores como em Ceron A. C (2014), Jensen (2013), Sanders (2013), Gaurav (2013) e Bermingham (2011), onde o Erro Médio Absoluto, a partir da contagem de tuítes, ficou entre 2% e 19%. Jungherr (2012) obteve 2.7% de MAE usando o último dia de dados e Tumasjan (2010) obteve 1.6% de MAE quando dados de 1 semana antes das eleições foram utilizados.

Tabela 4. MAE dos institutos de enquetes *offline* e do modelo baseado no volume de tuítes do primeiro e segundo turnos

1º Turno			2º Turno		
Insituto	**MAE**	**Acertou?**	**Insituto**	**MAE**	**Acertou?**
MDA Pesquisa	3,54%	Sim	Datafolha	0,36%	Sim
Vox Populi	4,15%	Sim	Contagem de tuítes	0,695%	Sim
Datafolha	4,21%	Sim	Ibope	1,36%	Sim
Ibope	4,54%	Sim	Vox Populi	2,36%	Sim
Contagem de tuítes	5,53%	Sim	MDA Pesquisa	3,54%	Não
Sensus	5,77%	Não	Sensus	5,77%	Não

Numa eleição cada eleitor vota apenas uma única vez em cada turno. Portanto, no modelo proposto para responder a segunda questão de pesquisa, a contagem de usuários será utilizada ao invés do volume de mensagens, porque independentemente da quantidade de mensagens publicadas por um usuário ele terá direito apenas a um único voto. Um usuário será considerado a favor de determinado candidato se o volume de mensagens publicadas por ele, contendo o nome daquele candidato for superior ao dos

outros presidenciáveis. Não foram considerados os usuários que não publicaram nenhuma mensagem sobre algum dos três presidenciáveis no período analisado. Usuários cujo volume de mensagens publicadas para cada candidato foi igual não foram considerados nesta análise. Assim como no primeiro modelo também foi utilizado o último dia que antecede cada um dos dois turnos como período de análise. Os resultados obtidos por este modelo são apresentados na Tabela 5.

Tabela 5. Resultado do modelo de predição baseado em contagem de usuários

	Parâmetros	Dilma	Aécio	Marina	MAE	Preveu corretamente?
1º turno	Volume de usuários	49,9%	24,8%	25,3%	6,65%	Não
	Resultado real	43,11%	34,78%	22,11%	-	-
2º turno	Volume de usuários	54,5%	45,5%	-	1,43%	Sim
	Resultado real	51,64%	48,36%	-	-	-

O segundo modelo previu corretamente apenas o resultado do segundo turno com MAE de 1,43%, diferentemente do primeiro modelo que previu corretamente os resultados de ambos os turnos. Considerando apenas os turnos onde ambos os modelos acertaram, o resultado obtido pelo primeiro foi melhor que o segundo. Este resultado se opõe aos experimentos conduzidos por Gaurav (2013) e Sang (2012). Nestas pesquisas, os autores conseguiram reduzir o MAE da eleição presidencial da Venezuela de 2% para 0,5% e na eleição presidencial do Equador de 19% para 3%, respectivamente, utilizando o segundo modelo ao invés do primeiro.

Para verificar a acurácia deste segundo modelo, 156 usuários da base de dados foram selecionados manualmente. Para participarem desta seleção os usuários deveriam ter publicado pelo menos uma mensagem sobre cada um dos candidatos no último dia que antecedeu cada um dos dois turnos da eleição e deveriam demonstrar apoio explícito ao candidato da preferência dele. Foram utilizados os critérios de Paula Filho (2015) para definição de apoio explícito aos candidatos (Tabela 6).

Tabela 6. Tipos de manifestações de apoio explícito a certo presidenciável

Tipos de apoio explícito	Exemplos
Mensagem de apoio a(o) candidata(o) e/ou partido no perfil do usuário	"Sou defensor ferrenho de Lula e Dilma"
	"... sou corretor de imoveis,atleticano,petista casado..."
Imagem de apoio a(o) candidata(o) no plano de fundo da conta do usuário	Imagem do ex-presidente Luiz Inácio Lula da Silva levantando a mão da presidenciável Dilma Rousseff
Imagem de apoio a(o) candidata(o) na foto do perfil do usuário	Foto do perfil do usuário com os dizeres "Dilma 13" na parte inferior da foto dele.
Mensagem de apoio a(o) candidata(o) nas mensagens públicas do usuário	"Voto útil, voto Aécio."
	"Boa noite Brasil lindo Marina 40 presidente"

No primeiro turno foram selecionados 106 usuários. Deste total, 36 manifestaram apoio explícito à candidata Dilma Rousseff, 32 a Aécio Neves e 38 a Marina Silva. Já no segundo turno foram selecionados 50 usuários. Deste total, 28 usuários demonstraram apoio explícito à candidata Dilma e 22 a Aécio. Para avaliar a acurácia deste modelo utilizou-se a medida de avaliação *Recall*. O resultado da análise para ambos os turnos são apresentados na Tabela 7.

Considerando o cenário político brasileiro, onde a margem de erro registrada pelo instituto *offline* com o melhor desempenho em ambos os turnos, o Datafolha, foi de

2 pontos percentuais para mais ou para menos (Eleições, 2014), utilizar um modelo cujos valores médios de *Recall* foram iguais a 25,6% e 24,35% no primeiro e segundo turnos, respectivamente, parece não ser uma escolha confiável. Mesmo apresentando um valor de *Recall* baixo no segundo turno o modelo conseguiu prever corretamente o resultado das eleições. Os valores próximos de *Recall* para ambos os candidatos no segundo turno parece ser uma justificativa plausível para o resultado positivo obtido pelo modelo.

Tabela 7. Acurácia do modelo de predição baseado em contagem de usuários

Presidenciável	Recall	
	1º turno	2º turno
Dilma Rousseff	80,6%	78,6%
Aécio Neves	68,8%	72,7%
Marina Silva	73,7%	-

Em Prasetyo (2014) resultados de várias pesquisas foram relacionados mostrando que a compreensão do sentimento de um *tweet* e a inclusão desta característica pode reduzir o erro da predição. Há vários métodos para realizar a análise de sentimentos. Segundo Prasetyo (2014) os mais utilizados pelos pesquisadores têm sido os métodos baseados em aprendizagem de máquina. Nós seguimos o estudo de (Ting, 2011) que comparou diversas abordagens de classificação, tais como Naive Bayes (NB), SVM, Árvore de Decisão e Redes Neurais. Os resultados obtidos por esta pesquisa mostraram que os dois melhores classificadores de texto foram o Naive Bayes e SVM, mas o Naive Bayes foi o mais amplamente utilizado por causa da simplicidade, menor tempo gasto na etapa de treinamento e classificação e o mais eficiente em uma variedade maior de domínios. Para responder a terceira questão de pesquisa deste trabalho, utilizou-se este algoritmo para criação do modelo de predição.

Para a etapa de treinamento 18.674 amostras foram rotuladas utilizando o método proposto por Paula Filho (2015). Seguindo o modelo proposto por Prasetyo (2014) as mensagens foram rotuladas em positivo e negativo (Tabela 8). Foram selecionadas mensagens de todos os usuários que publicaram *tweets* no último dia antes da votação do segundo turno. No total, foram selecionadas mensagens de 40.587 usuários distintos. Como o volume de mensagens publicado no último dia pela maioria dos usuários foi baixo, em alguns casos 1 ou 2 mensagens, optou-se por aumentar o tempo de coleta para 7 dias para o mesmo conjunto de usuários. Foram excluídas das mensagens símbolos de retuítes, menções a usuários e endereços eletrônicos. A biblioteca WEKA[6] para Java foi utilizada para avaliar a classificação do algoritmo. A Tabela 9 apresenta os valores de *Precision*, *Recall* e *F-Measure* e a Tabela 10 a matriz de confusão obtida ao término da fase de treinamento.

Segundo Wiebe (2006) e Golden (2011), a capacidade humana de avaliação correta da subjetividade de um texto varia de 72% a 85%, respectivamente. Portanto, pode-se concluir que os resultados são satisfatórios.

Após a etapa de treinamento o modelo foi salvo para ser utilizado na classificação de novas instâncias. Na fase de classificação a biblioteca WEKA para Java foi novamente utilizada. Foram classificadas 145.959 mensagens de 14.790 usuários.

[6] WEKA (Waikato Environment for Knowledge Analysis) - http://www.cs.waikato.ac.nz/ml/weka/

Deste total, 61.483 mensagens diziam respeito à candidata Dilma e 84.476 ao candidato Aécio Neves.

Tabela 8. Volume de mensagens rotuladas

	Dilma Rousseff	Aécio Neves
Positivo	4.587	4.693
Negativo	4.441	4.953
Total	9.028	9.646

Tabela 9. Detalhamento das acurácias por classe/candidato

Classe	Candidato	Precision	Recall	F-Measure
Positivo	Dilma	0.781	0.767	0.774
	Aécio	0.821	0.78	0.8
Negativo	Dilma	0.763	0.778	0.77
	Aécio	0.801	0.839	0.82

Tabela 10. Matriz de confusão

Categoria	Candidato	Classificação como positivo	Classificação como negativo
Positivo	Dilma	3517	1070
	Aécio	3659	1034
Negativo	Dilma	988	3453
	Aécio	796	4157

Para cada usuário foi computado o volume de mensagens positivas e negativas a respeito de cada candidato. Os três possíveis resultados para este modelo baseado na polaridade foram: (a) Dilma - quando o volume de mensagens positivo para esta candidata fosse superior ao negativo e o volume de mensagens negativo for superior ao positivo para o candidato Aécio, (b) Aécio - quando o volume de mensagens positivo para este candidato for superior ao negativo e o volume de mensagens negativo for superior ao positivo para a candidata Dilma e (c) Descarte de mensagens, quando qualquer situação for diferente das duas anteriores. O resultado deste modelo é apresentado na Tabela 11.

Tabela 11. Resultado do modelo de predição baseado na análise de sentimento

Presidenciáveis/ Resultado do modelo de predição	Modelo - 2º turno	Resultado real (votos válidos) 2º turno
Dilma	46,3%	51,64%
Aécio	53,7%	48,36%
Preveu correto o ganhador?	Não	-
MAE	2,67%	-

O terceiro modelo de predição, baseado na análise de sentimentos, não conseguiu acertar o resultado das eleições no segundo turno, diferentemente dos outros dois modelos anteriores baseado em contagem. Para verificar a precisão deste modelo 20 usuários que manifestaram apoio explícito a cada um dos dois candidatos foram selecionados entre os 14.790 usuários utilizados na etapa de classificação. De todos os usuários selecionados o modelo conseguiu acertar 100% das escolhas dos usuários. Apesar do modelo não ter acertado o resultado das eleições ele foi mais eficiente do que o modelo baseado na contagem de usuários.

6. Conclusões e trabalhos futuros

A contagem de tuítes de um dia antes dos dois turnos das eleições para presidente no Brasil foi capaz de prever corretamente os resultados de ambos os turnos. A eliminação de tuítes repetidos pareceu não ser uma boa estratégia para melhorar o MAE. No primeiro turno, a remoção de tuítes repetidos produziu uma predição errada e no segundo turno piorou o valor do MAE. O terceiro modelo apresentado, baseado na análise de sentimento das mensagens, apesar de não ter acertado o resultado das eleições do segundo turno, pareceu ser uma boa opção, pois foi o que apresentou melhor precisão. Acredita-se que ele só não apresentou uma eficiência melhor, pois algumas possibilidades do modelo não foram contempladas no modelo, como por exemplo, situações em caso de empate (negativo/negativo, positivo/positivo e empate/empate) entre o volume de mensagens positivas e negativas entre os candidatos.

Como proposta de trabalhos futuros, pretende-se propor soluções para as etapas de Filtragem de Dados e Redução de viés de dados. Conforme apresentado anteriormente, seção 2, resultados de diversas pesquisas tem conseguido melhorar os resultados dos modelos através destas técnicas. Pretende-se também aprimorar o modelo baseado em tuítes, considerando-se outras possibilidades de combinações entre os resultados obtidos pelo volume das mensagens. Além disso, pretende-se desenvolver um modelo capaz de analisar situações onde três ou mais candidatos fazem parte da corrida presendencial.

7. Referências

Bakker, T. P. (2011). Good news for the future? Young people, Internet use, and political participation. Communication Research.

Beauchamp, N. (2013). Predicting and interpolating state-level polling using twitter textual data. Meeting on automated text analysis, London School of Economics.

Cameron, M. P. (2013). Can Social Media Predict Election Results? Evidence from New Zealand. No. 13/08.

Ceron, A. C. (2014). Using Sentiment Analysis to Monitor Electoral Campaigns: Method Matters—Evidence From the United States and Italy. Social Science Computer Review.

Choy, M. C. (2011). A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction. arXiv preprint arXiv:1108.5520.

Choy, M. C. (2012). US Presidential Election 2012 Prediction using Census Corrected Twitter Model. arXiv preprint arXiv:1211.0938.

Chung, J. E. (2011). Can collective sentiment expressed on twitter predict political elections? AAAI.

Eleições 2014. Pesquisa eleitoral para presidente. Disponível em: http://www.eleicoes2014.com.br/pesquisa-eleitoral-para-presidente/. Acesso em: 10 abr. 2014.

Fink, C. B. (2013). Twitter, Public Opinion, and the 2011 Nigerian Presidential Election. 2013 International Conference on Social Computing (SocialCom). IEEE., 311-320.

Fumagalli, L. &. (2011). The total survey error paradigm and pre-election polls: The case of the 2006 Italian general elections. ISER Working Paper Series. 2011-29.

Gaurav, M. S. (2013). Leveraging candidate popularity on Twitter to predict election outcome. Proceedings of the 7th Workshop on Social Network Mining and Analysis. ACM., 7.

Gayo Avello, D. (2011). Don't turn social media into another'Literary Digest'poll. Communications of the ACM, 54(10), 121-128.

Hillygus, D. S. (2011). The evolution of election polling in the United States. Public opinion quarterly, 75(5),, 962-981.

Jungherr, A. J. (2012). Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, to, sander, pg, & welpe, im "predicting elections with twitter: What 140 characters reveal about political sentiment". Social Science Computer Review, 30(2),, 229-234.

Lewis-Beck, M. S. (2005). Election forecasting: principles and practice. The British Journal of Politics & International Relations, 7(2),, 145-164.

Makazhanov, A. R. (2014). Predicting political preference of Twitter users. Social Network Analysis and Mining, 1-15.

Mejova, Y. S. (2013). GOP primary season on twitter: popular political sentiment in social media. Proceedings of the sixth ACM international conference on Web search and data mining, 517-526.

Mislove, A. L. (2011). Understanding the Demographics of Twitter Users. ICWSM, 11, 5th.

Nooralahzadeh, F. A. (2013). 2012 Presidential Elections on Twitter--An Analysis of How the US and French Election were Reflected in Tweets. Control Systems and Computer Science (CSCS), 2013 19th International Conference on, 240-246.

O'Connor, B. B. (2010). From tweets to polls: Linking text sentiment to public opinion time series. ICWSM, 11, 122-129.

Paula Filho, W.; GARCIA, A. C. B. RoTuEl: A Semi-Automated Method For Labeling Political. In: Doctoral Consortium at International Joint Conference on Artificial Intelligence (IJCAI-15), 2015, Buenos Aires, Argentina. Proceedings of the Twenty-Four IJCAI.

Prasetyo, N. D. (2014). Tweet-Based Election Prediction (Doctoral dissertation, TU Delft, Delft University of Technology).

Sang, E. T. (2012). Predicting the 2011 dutch senate election results with twitter. the Workshop on Semantic Analysis in Social Media (pp. 53-60). Association for Computational Linguistics.

Tumasjan, A. S. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. ICWSM, 10, 178-185.

Wong, F. T. (2013). Media, pundits and the us presidential election: Quantifying political leanings from tweets. In Proceedings of the International Conference on Weblogs and Social Media.