PMKT
REVISTA BRASILEIRA DE PESQUISAS DE MARKETING, OPINIÃO E MÍDIA

Editor-Chefe: Fauze Najib Mattar | Editor Adjunto: Claudio Alberto de Moraes
Sistema de avaliação: Triple Blind Review
Publicação: Associação Brasileira de Empresas de Pesquisa (ABEP)
Idiomas: Português e Inglês

Índice de sentimento de candidatos e intenção de voto. Podem esses indicadores coexistirem?

Index of candidates and voting. Can these indicators coexist?

Rodrigo Otávio de Araújo Ribeiro*, Reinaldo Gomes Morais
IBOPE DTM, Rio de Janeiro, RJ, Brasil

Patrícia Pavanelli, Bruna Suzzara Bueno de Miranda
IBOPE Inteligência, São Paulo, SP, Brasil

RESUMO

Este estudo tem como objetivo avaliar a relação existente entre o índice de sentimento dos principais candidatos no Twitter, durante a campanha eleitoral para presidência de 2014 e a intenção de voto dos brasileiros, captada pelas pesquisas realizadas pelo IBOPE no mesmo período. A utilização de mais de uma fonte de informação em análises de dados constitui um dos alicerces do Big Data. Foi verificado que índices relacionados ao volume de postagens possuem maior correlação com os resultados das pesquisas realizadas do que os que se baseiam na avaliação do sentimento. Uma análise de tópicos complementar também foi realizada em períodos imediatamente anteriores aos turnos da eleição, possibilitando a rápida identificação dos assuntos postados no Twitter sobre as candidaturas.

PALAVRAS-CHAVE: Eleições presidenciais; Análise de sentimento; Twitter.

ABSTRACT

This study aims to evaluate the relationship between the sentiment index of the leading candidates on Twitter during the election campaign for the presidency in 2014 and the voting intention of Brazilians captured by IBOPE surveys on the same period. The use of more than one source of information in data analysis is one of the foundations of Big Data. It was found that metrics related to the volume of posts have higher correlation with the results of surveys than those based on sentiment analysis. A Topic Analysis was also performed considering the periods immediately prior to the days of the elections, enabling a faster identification of subjects posted on Twitter about the campaign.

KEYWORDS: *Presidential Elections; Sentiment Analysis; Twitter.*

Submissão: 19 maio 2016
Aprovação: 23 agosto 2017

***Rodrigo Otávio de Araújo Ribeiro**
Doutor em Engenharia de Produção pela Universidade Federal Fluminense (UFF). Diretor de Inteligência de Marketing no IBOPE DTM. Professor no Instituto de Matemática e Estatística (IME) da Universidade do Estado do Rio de Janeiro (UERJ).
(CEP 22270-000 - Botafogo, Rio de Janeiro, RJ, Brasil).
E-mail: rodrigo.ribeiro@ibopedtm.com
Endereço: Rua Voluntários da Pátria 45, sala 1308 - 22270-000 - Botafogo, Rio de Janeiro, RJ, Brasil.

Reinaldo Gomes Morais
Mestrando em Engenharia Eletrônica pela Universidade do Estado do Rio de Janeiro. Analista Pleno de Inteligência de Marketing no IBOPE DTM.
E-mail: reinaldo.gomes@ibopedtm.com

Patrícia Pavanelli
Pós-Graduada em Gestão Pública pela Fundação Escola de Sociologia e Política de São Paulo. Diretora de contas, opinião pública, política e comunicação do IBOPE Inteligência.
E-mail: patricia.pavanelli@ibopeinteligencia.com

Bruna Suzzara Bueno de Miranda
Graduada em Estatística pela Universidade Estadual de Campinas (UNICAMP). Coordenadora de Estatística no IBOPE Inteligência.
E-mail: bruna.suzzara@ibopeinteligencia.com

1 INTRODUÇÃO

Este estudo teve como objetivo avaliar a relação existente entre o índice de sentimento de candidatos no Twitter, durante a campanha eleitoral para presidência de 2014 e a intenção de voto dos principais candidatos, captada pelas pesquisas realizadas pelo IBOPE no mesmo período. Foi observada uma maior relevância do volume de postagens em relação ao sentimento dos comentários na correlação gerada com as taxas de intenção de votos oriundas das pesquisas. Este resultado foi surpreendente, pois era esperada uma alta correlação entre o índice de sentimento e a intenção de votos. Talvez, este fenômeno pudesse ser explicado pela falta de aderência do modelo se sentimento. Contudo, isso não se verificou, uma vez que a performance do modelo em questão foi satisfatória.

Nos últimos anos, muitos estudos com base em informações advindas de redes sociais têm sido desenvolvidos. Contudo, são raros os estudos que buscam o comparativo entre resultados utilizando redes sociais e pesquisas quantitativas tradicionais. A utilização de mais de uma fonte de informação em análises de dados constitui um dos alicerces do *Big Data*, no qual não apenas o volume e velocidade da informação são importantes, mas a variedade também exerce um papel fundamental para uma visão mais clara acerca do assunto de interesse.

Nos Estados Unidos, O'Connor, Balasubramanyan, Routledge e Smith (2010) realizaram um estudo com características próximas. Porém, os mesmos autores obtiveram uma correlação positiva entre métricas de sentimento captadas via Twitter e resultado de pesquisas de aprovação no governo do Presidente Obama.

Os resultados deste estudo servem de auxílio para profissionais e empresas de pesquisa que atuem no setor político, no sentido de possibilitar uma avaliação mais rica acerca do cenário eleitoral brasileiro em épocas de eleição. Foi possível entender as limitações e benefícios das informações geradas por meio da análise de dados da rede social Twitter no contexto eleitoral, assim como a correlação existente com os indicadores eleitorais tradicionais.

Os resultados foram obtidos com base nas informações das pesquisas quantitativas de intenção de voto realizadas pelo IBOPE e no monitoramento dos principais candidatos à presidência da república nas eleições de 2014: Dilma Rousseff, Aécio Neves, Eduardo Campos e Marina Silva (Marina Silva substituiu Eduardo Campos após o falecimento deste).

A análise de sentimento dos candidatos no Twitter foi feita com base na metodologia desenvolvida pelo IBOPE DTM. O algoritmo realiza a leitura da postagem contendo o nome dos candidatos, para classificá-la em positiva, neutra ou negativa.

O IBOPE realiza pesquisas de intenção de voto de abrangência nacional com amostras proporcionais ao número de eleitores de cada região do país. Desta maneira, a intenção de voto a determinado candidato é calculada pela proporção de pessoas que declararam que votariam no mesmo, caso a eleição fosse na data de realização da pesquisa.

Vale ressaltar que o IBOPE DTM possui todas as postagens feitas no Twitter, realizadas em língua portuguesa com os nomes dos candidatos, no período estudado. As postagens foram capturadas por meio da ferramenta GNIP. A análise dos principais tópicos relacionados aos candidatos que tiveram destaque no Twitter, em cada um dos momentos da campanha nos quais ocorreram as pesquisas, também foi realizada.

2 PESQUISAS ELEITORAIS

2.1 Objetivos e histórico no Brasil

A partir da redemocratização do país nos anos de 1980 que permitiu aos brasileiros o exercício do voto para a escolha de seus governantes depois de anos de Ditadura Militar, o Brasil realizou sete eleições presidenciais, sendo a última, em 2014. Assim como aconteceu em outros países democráticos, a realização e divulgação de pesquisas quantitativas de intenção de voto tornou-se parte do contexto das eleições do país.

A pesquisa de opinião é uma fonte de informação a respeito do pensamento geral de uma população sobre os temas sociais e políticos de um país.

Nesse contexto, as pesquisas de intenção de voto (políticas, eleitorais) são ferramentas importantes e eficazes para o conhecimento da opinião e do comportamento dos eleitores, e possibilitam entender como se manifesta a intenção de voto dos indivíduos dentro do grupo social.

De modo geral, as pesquisas eleitorais representam sempre um instante da realidade, um retrato do momento. Como a fotografia, o resultado de uma pesquisa é uma imagem inerte de algo que está em constante movimento: a opinião. As pesquisas não predizem o futuro, elas indicam tendências que podem ser alteradas se algo interferir na realidade medida, fazendo com que mude a opinião pública. A opinião pública, neste artigo, é compreendida como o resultado de respostas a perguntas de entrevistas (Lane & Sears, 1964; Converse, 1987; Price, 1992; Boyte, 1995; Worcester, 1997; SamPedro, 2000; Sartori, 2002), diferentemente dos conceitos que a mostram como um processo deliberativo promovido por cidadãos informados e participantes ativos da vida democrática, conforme propõem Speier (1950), Habermas (1998) e Bourdieu (1973).

Segundo Marcia Cavallari, Diretora Executiva do IBOPE Inteligência, em entrevista ao site Congresso em Foco:

> Pesquisa não é infalível, não dita a última palavra. É uma informação a mais que o eleitor tem em meio a tantas outras para a tomada de decisão. Cada vez mais as pesquisas têm de ser interpretadas como diagnóstico do momento. São uma fotografia do momento. A sequência dessas fotografias vai montando um filme, com as tendências. Quando a gente divulga a pesquisa da véspera, não significa que o processo de consolidação de voto se congela, que ninguém muda mais. (Cavallari, 2016)

Atualmente há diversas propostas de leis para controlar a realização das pesquisas eleitorais e sua divulgação pela mídia. Desde 1997, o Tribunal Superior Eleitoral regula a divulgação dos resultados de pesquisas eleitorais, obrigando que, qualquer levantamento a ser divulgado, seja registrado para que sua publicação ocorra após o prazo estipulado pelo órgão. Ao longo do ano, o TSE registrou 2.411 pesquisas eleitorais (Gramacho, 2014).

Pioneiro neste tipo de pesquisa no Brasil, tendo iniciado a realização e a divulgação de levantamentos eleitorais em 1945, o IBOPE acompanhou as sete eleições do período pós-ditadura e, por isso, pode ser considerado um dos mais importantes e um dos maiores conhecedores do comportamento do eleitor brasileiro do país.

Há tempos, o IBOPE é o instituto de pesquisa do país, responsável por medir e divulgar o maior volume de levantamentos eleitorais no Brasil, sendo grande parte deles divulgados no mais expressivo veículo de comunicação brasileiro. Com esta importância e alcance, os resultados são repercutidos por todos os demais veículos e comentados tanto pela crítica especializada quanto pela população em geral.

2.2 Processo de construção da amostra

As amostras nacionais realizadas pelo IBOPE Inteligência têm como objetivo refletir a opinião do eleitorado brasileiro que votou nas últimas eleições (votantes).

Ao planejar esse tipo de estudo, esbarra-se na limitação/desatualização das informações que existem nos cadastros do TSE. Essas informações refletem, em sua maioria, as características dos eleitores no momento que obtém seus títulos de eleitores. Informações como idade e grau de instrução não são atualizadas nessas bases oficiais.

Buscando atualizar o perfil atual do eleitorado, agregam-se aos dados do TSE, estimativas populacionais realizadas pelo IBOPE Inteligência baseadas em dados oficias (Censo e PNAD mais atuais), além de estudos internos. Essas informações auxiliam no momento da elaboração das cotas da amostra.

O universo de votantes é estratificado por estado, com exceção dos estados do Acre, Amapá e Roraima que, juntos, constituem apenas um estrato. Uma vez que o Estado possui Região Metropolitana, o seu universo é estratificado em Região Metropolitana e Interior. Em seguida, é selecionada uma amostra de conglomerados em três estágios:

- No primeiro estágio, os municípios são selecionados probabilisticamente por meio do método Probabilidade Proporcional ao Tamanho (PPT), tomando os eleitores que votaram nas últimas eleições (votantes) como base para tal seleção;
- No segundo estágio, são selecionados os conglomerados: setores censitários, com PPT sistemático. A medida de tamanho é o número de votantes dos setores;
- No terceiro estágio, é selecionado, em cada conglomerado, um número fixo de votantes segundo cotas de: sexo, idade, instrução e condição de atividade.

3 TWITTER

O Twitter foi criado em 2006 pelos sócios Jack Dorsey, Evan Williams, Biz Stone e Noah Glass, em San Francisco – EUA. O serviço é uma rede social que permite aos usuários postarem e lerem *tweets*, que nada mais são, que mensagens de até 140 caracteres. Seu acesso pode ser feito diretamente em algum *browser* de internet, por aplicativos no celular e, em alguns países, as postagens podem ser feitas por meio de SMS. A ideia rapidamente se espalhou e ganhou popularidade no mundo todo: em 2012, eram mais de 500 milhões de usuários registrados que postavam 340 milhões de *tweets* por dia (Lunden, 2012).

Uma vez cadastrado, o usuário define um endereço no site que ainda não está sendo utilizado; a partir de então, ele será sempre conhecido por esse endereço precedido do símbolo @ pelos outros usuários.

Definido esse endereço e cadastrada a conta, o usuário poderá seguir ou ser seguido por outras contas. Isso significa que, cada vez que usuários seguidos postam algo, a mensagem aparece diretamente na sua página (também chamada de *timeline*). Por *default*, *tweets* são visíveis publicamente, no entanto, é possível restringir a visualização das mensagens para apenas seus seguidores. Outra possibilidade de mensagem é repostar o que já foi postado por alguém, prática também conhecida como *retweet*, e que é caracterizada pela sigla RT. O objetivo, nesse caso, é o usuário repassar esse determinado texto para todos que o seguem (Strachan, 2009).

Quando uma postagem é feita em cima de um tópico específico, o usuário pode fazer uso de uma técnica chamada *hashtag* – frases ou palavras que começam com o símbolo # (Strachan, 2009). Da mesma forma, se o interesse for visualizar apenas mensagens daquele tópico, uma busca pode ser feita utilizando o mesmo termo em *hashtag*.

4 CAMPANHA PRESIDENCIAL DE 2014

A eleição de 2014, a sétima disputa presidencial brasileira desde a redemocratização nos anos de 1980, foi a mais acirrada que o país já teve. Dilma Rousseff foi reeleita com 51,6% dos votos válidos, em segundo turno, o que representa a vitória mais apertada desde o fim da Ditatura Militar, evidenciando o latente desejo dos eleitores pela mudança na condução do governo brasileiro.

As pesquisas realizadas pelo IBOPE indicavam que, em 2014, cerca de 70% dos eleitores desejavam que o próximo presidente mudasse totalmente os programas e medidas do Governo Federal ou mantivesse apenas alguns deles. Este índice só é inferior ao observado nos levantamentos de 2002, ano da eleição de Lula. Este cenário foi bem diferente na eleição de 2010, no qual o desejo de continuidade prevalecia para seis em cada dez eleitores.

Após anos de crescimento econômico durante os governos Lula, o primeiro mandato de Dilma Rousseff reflete mais um período de estagnação do que de avanços, marcado pela combinação do aumento de preços, da queda do poder aquisitivo, do endividamento das famílias e de reajustes de

tarifas públicas. O descontentamento geral de milhões de brasileiros foi marcado pelas manifestações de julho de 2013 e acentuado pelos gastos nas obras para a realização da Copa do Mundo de Futebol no Brasil, ocorrida em julho de 2014.

Após as manifestações ocorridas em 2013 contra o aumento das tarifas de ônibus e por melhorias nos serviços públicos, a realização, em 2014, da Copa do Mundo de Futebol, e o baixo desempenho da economia brasileira, o clima eleitoral da disputa presidencial era, por parte dos eleitores, de pouco interesse e de intensa desilusão em relação à política. Tais aspectos podem ser observados nos resultados da Figura 1, que mostra um aumento significativo dos brasileiros que declaram que não iriam votar se o voto não fosse obrigatório (passam de 35% em 2010 para 50% dos eleitores em 2014).

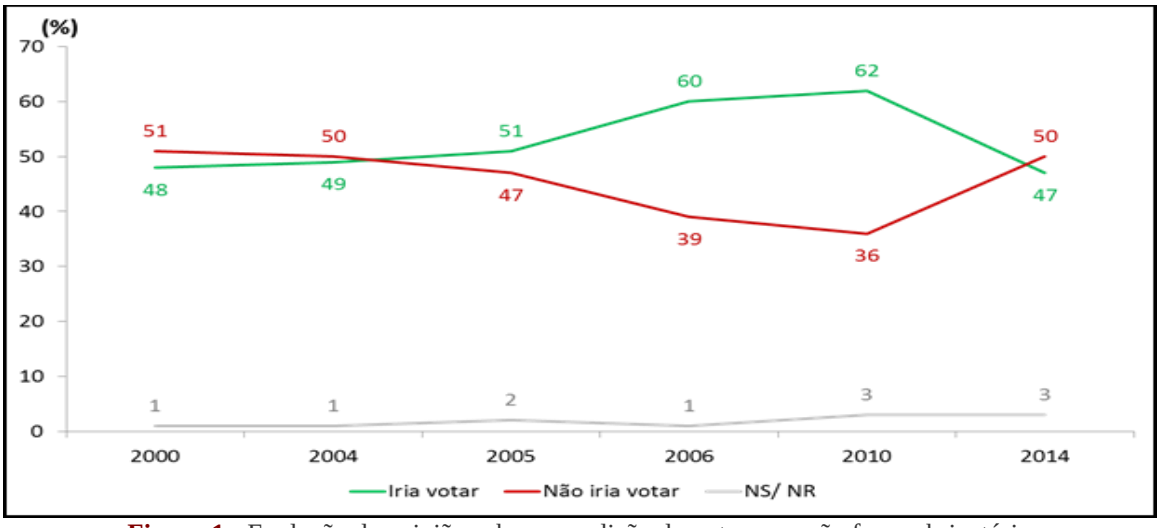

Figura 1 - Evolução da opinião sobre a condição do voto, caso não fosse obrigatório

Com a oficialização das candidaturas[1], a campanha se consolida com três principais candidatos: a Presidente Dilma Rousseff pelo Partido dos Trabalhadores (PT), o Senador Aécio Neves pelo Partido da Social da Democracia Brasileira (PSDB) e o ex-Governador de Pernambuco Eduardo Campos pelo Partido Socialista Brasileiro (PSB). Entretanto, em 13 de agosto, a campanha eleitoral registrou uma grande tragédia: o falecimento do candidato do PSB, Eduardo Campos. O candidato estava a bordo de seu avião de campanha, juntamente com quatro assessores e dois pilotos que também perderam a vida. O acidente comoveu o país e transformou a dinâmica da disputa eleitoral. Em terceiro lugar nas pesquisas de intenção de voto, Campos foi substituído por Marina Silva, tendo como vice, o deputado Beto Albuquerque do PSB do Rio Grande do Sul.

Como pode ser observado na Figura 2, Dilma Rousseff esteve na liderança em todas os levantamentos feitos pelo IBOPE no primeiro turno. A partir do momento que se tornou candidata, Marina Silva que, embora tenha terminado a eleição em 3º lugar, apresentou um bom desempenho até a véspera do pleito. Ao ser atacada pelos adversários por suas declarações e propostas contraditórias foi perdendo intenções de voto e não conseguiu se sustentar até o fim. Aécio Neves que, embora não tenha conseguido captar os votos de Marina na mesma velocidade em que ela caía, reassumiu a segunda posição, após o último debate da campanha eleitoral, às vésperas da eleição. O resultado oficial do primeiro turno estabeleceu, mais uma vez, a disputa ao posto pela Presidência da República em segundo turno entre PT e PSDB.

[1] Foram registradas, no Tribunal Superior Eleitoral, 11 candidaturas para o cargo de Presidente da República (Tribunal Superior Eleitoral [TSE], 2016).

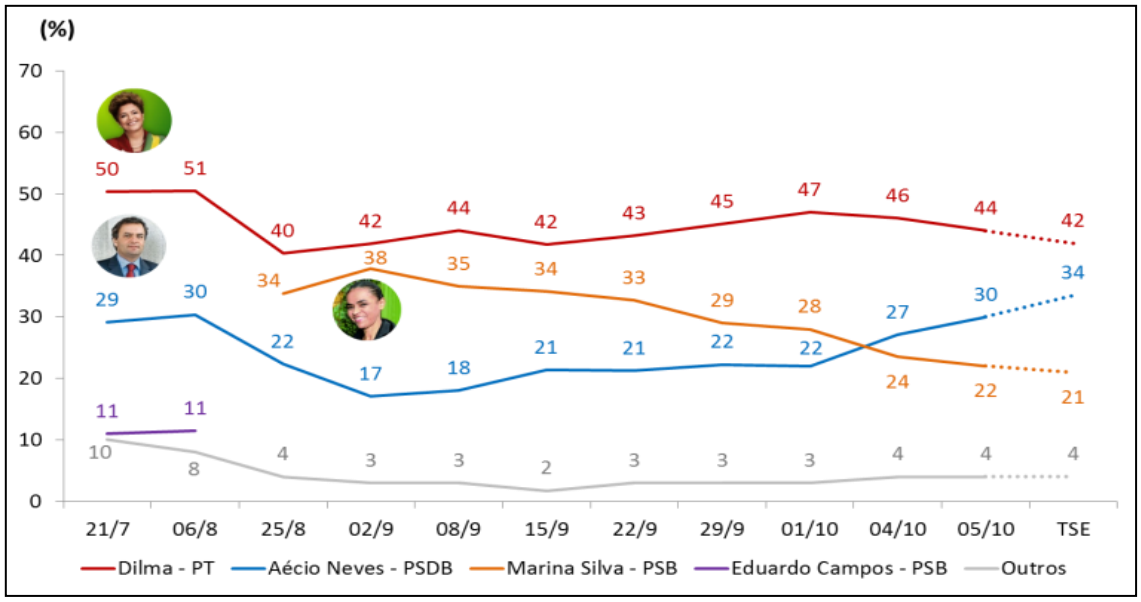

Figura 2 - Evolução das intenções de voto para presidente e Resultado Oficial do 1º Turno das Eleições Presidenciais – Votos válidos

Foi no segundo turno da campanha que o acirramento da disputa entre os candidatos se potencializou. As primeiras pesquisas apresentavam Aécio Neves numericamente à frente de Dilma, mas a cinco dias da eleição a situação se inverteu. A Presidente Dilma Rousseff ultrapassou o candidato do PSDB e ganhou a eleição com 51,6% dos votos válidos, conforme mencionado anteriormente, a vitória mais apertada desde a redemocratização do país.

5 METODOLOGIA ANALÍTICA

A metodologia analítica aplicada neste artigo consiste na execução de quatro passos: o primeiro refere-se à análise da qualidade de ajuste do modelo de sentimento (em relação ao candidato) desenvolvido; o segundo busca avaliar o comportamento geral dos usuários e o perfil dos mesmos quanto a utilização do Twitter para realizar postagens sobre política, sendo feita também uma breve descrição do perfil do eleitor brasileiro; o terceiro é relativo a análise da correlação entre intenção de voto e índices oriundos do Twitter; no quarto passo, realiza-se uma análise semântica que se baseia na utilização de técnicas de *Text Mining* para identificação dos tópicos mais pertinentes dentro do ambiente de conversa das eleições presidenciais.

5.1 Modelo de sentimento (em relação aos candidatos)

5.1.1 *Text Mining*

A Mineração de Textos, também conhecida como *Text Mining*, é o processo de extração de informação útil ou conhecimento de documentos de textos estruturados ou não (Barion & Lago, 2008). No contexto desse artigo, essa técnica será aplicada para identificar padrões de comentários e opiniões emitidas por usuários do Twitter sobre os candidatos à presidência da eleição Nacional de 2014.

O primeiro passo da mineração é a indexação, processo que armazena uma estrutura de índices a partir das palavras dos textos e viabiliza a pesquisa de documentos por meio de todos os termos contidos ali (Salton & McGill, 1983). Segundo Barion e Lago (2008), algumas etapas importantes para uma análise de *Text Mining* devem ser seguidas e são:

- Análise Léxica: converte uma sequência de caracteres numa sequência de palavras que serão candidatas a termos do índice. Nesta fase, são separados o alfabeto de entrada em caracteres de palavras e separadores de palavras;
- Remoção de *stopwords*: remove um conjunto de palavras que aparecem com frequência em textos, mas não possuem valor semântico, tais como: preposições, artigos e conjunções. Essa fase é de extrema importância, pois diminui a base a ser indexada e facilita a mineração;
- *Stemming:* remove todas as variações de palavras, deixando apenas a raiz de cada uma, por exemplo, a palavra "amamos" passa a se identificar como a raiz "ama";
- Seleção dos termos-índice: determina quais palavras ou radicais serão utilizados como elementos de indexação. Estas palavras são selecionadas de acordo com o peso atribuído a elas;
- *Bag of words (BOW)*: consiste em uma matriz na qual cada termo diferente presente na coleção de documentos é indexado. A partir desta indexação, cada documento pode ser representado por um vetor 1 x *n*, onde *n* é o número total de termos, cada entrada desse vetor será o número de vezes que os termos aparecem nesse documento (Sivic, 2009);
- Determinação dos pesos: o preenchimento da matriz *BOW* é feito com base em métricas que ponderam a frequência dos termos nos documentos e na coleção total (conjunto de todos os documentos). A métrica mais comumente utilizada para esta finalidade é denominada *tf-idf (term frequency - inverse document frequency).*

5.1.2 Identificação do sentimento

O Índice de Sentimento serve como termômetro da polaridade das mensagens postadas sobre política no Twitter. Trata-se de uma métrica comparativa que possibilita a avaliação de candidatos em determinado momento, assim como sua evolução pelo tempo.

A elaboração do algoritmo ocorreu em quatro etapas:

- Etapa 1 - Desenvolvimento de metodologia de polarização feita por especialistas em língua portuguesa;
- Etapa 2 - Polarização manual em amostra seguindo a metodologia desenvolvida na etapa 1;
- Etapa 3 - Desenvolvimento do algoritmo para polarização automática de postagens sobre política com base na amostra polarizada da etapa 2;
- Etapa 4 - Implementação, avaliação do erro e recalibragem com base em novas amostras coletadas.

O método desenvolvido combina o conhecimento humano com sofisticados modelos estatísticos para avaliação da polaridade de determinada postagem. Modelos de análise de sentimento apresentam as seguintes características:

- Capacidade de lidar com grandes volumes de dados com alta velocidade de resposta;
- Garantia de mesmo critério de avaliação e mesma regra aplicada a todas as postagens.

Já a marcação manual da polaridade feita por agentes humanos apresenta as seguintes características:

- Quando bem treinado, o classificador consegue captar as nuances interpretativas do texto;
- Incapacidade de lidar com grandes volumes de dados em alta velocidade;
- Dificuldade de treinamento de equipe de polarizadores em manter um padrão homogêneo de classificação nas marcações.

Estudo realizado pela TOPSY nos EUA, com três classificadores polarizando uma amostra de 10 mil postagens relacionadas à política, obtiveram concordância 73% (Gosh, 2016). Foi realizado um estudo semelhante no IBOPE DTM comparando-se marcações das mesmas postagens por diferentes pessoas e foi encontrada uma concordância de 59% nas classificações, considerando-se a classificação dos especialistas com a de um classificador de nível superior. A importância da consistência da marcação manual que alimentará o modelo é um aspecto essencial e mais crítico da metodologia de análise de sentimento desenvolvida.

Algoritmos de análise de sentimento atribuem diferentes graus de importância às palavras que aparecem nas postagens, símbolos ou expressões presentes de acordo com sua frequência de ocorrência nas mensagens avaliadas. Estes graus de importância ou pesos, são obtidos durante o processo de modelagem estatística que possui como objetivo, a elaboração de regras para classificação correta das mensagens polarizadas manualmente.

O algoritmo desenvolvido, considera para sua calibração, apenas postagens classificadas por especialistas relacionadas aos candidatos, sendo assim, os termos linguísticos têm sua importância medida considerando a conotação correta. Para exemplificar este aspecto, basta imaginar as diferentes polaridades que podem ser atribuídas a palavra "encoberto" nas seguintes frases: "foi evidente que o governador havia encoberto o esquema de lavagem de dinheiro"; "mesmo com o céu encoberto nesta manhã de quinta-feira, não existe possibilidade de chuvas fortes".

Na primeira frase, pode-se observar que a palavra "encoberto" associada ao assunto "política", apresenta uma polaridade negativa. Já quando está relacionada ao assunto "meteorologia", pode apresentar polaridade positiva. Logo, se um modelo estatístico fosse elaborado para marcar a polaridade no ambiente de conversas "políticas" fosse ajustado com base em uma amostra de postagens relacionadas a "meteorologia", não geraria informações consistentes. Por este motivo, é essencial que os modelos de análise de sentimento sejam calibrados levando-se em conta amostras associadas ao contexto que se deseja modelar.

O algoritmo desenvolvido marca um *score* de sentimento em todas as postagens realizadas sobre os candidatos considerados. O *score* de sentimento é uma medida que varia de 0 a 100, quanto mais próximo de 0, mais negativa é a mensagem e, quanto mais próximo de 100, mais positiva. Quando o *score* assume valores acima de 60, a postagem é considerada positiva, entre 40 e 60, neutra e, abaixo de 40, negativa.

5.1.3 *Random Forest*

O método de classificação de sentimento desenvolvido utiliza como classificador o modelo de *Random Forest*, desenvolvido por Breiman (2001). Consiste na utilização conjunta de múltiplas árvores de decisão aleatórias (*Random Trees*) visando a melhoria da classificação. Conforme o autor, dentre as características positivas deste método se destacam:

- Excelente acurácia, quando comparado a outros algoritmos de classificação;
- Boa performance em grandes bases de dados;
- Capacidade de lidar com milhares de variáveis preditoras, sem a necessidade de deleção;
- Capacidade de dizer quais variáveis são mais importantes para classificação;
- Geração de uma estimativa não viciada do erro por meio do processo de construção das florestas;
- Proposta eficiente de estimação com informações faltantes, quando este problema existe na base de dados a ser analisada;
- Possibilidade de balancear o erro em bases de dados desbalanceadas;
- Possibilidade de utilizar as florestas para previsão de amostras futuras;
- Cálculo das proximidades entre pares de casos, informação esta que pode ser usada na detecção de *outliers* ou na clusterização dos dados.

Para verificar a acurácia da marcação feita pelo modelo de análise de sentimento, amostras de postagens sobre os candidatos foram extraídas (durante o período das eleições) para serem polarizadas manualmente por profissionais treinados na metodologia de polarização manual desenvolvida pelos especialistas em língua portuguesa. O trabalho deles consistia em avaliar se a marcação do modelo automático estava ou não certa.

5.2 Avaliação da evolução de postagens e perfil dos usuários que interagem com a campanha pelo Twitter

Na primeira etapa analítica, buscou-se avaliar as principais métricas agregadas, presentes no trabalho, agrupadas no tempo. As mais importantes foram as seguintes:

- Quantidade total de postagens: avalia o número total de postagens realizadas por intervalo de tempo;
- Quantidade de postagens por usuário;
- Penetração de postagens: percentual de postagens com dada característica de interesse.

A análise da quantidade total de postagens torna possível a avaliação da intensidade dos impactos ocorridos durante o período observado. Já a penetração de postagens, avalia o peso da existência de determinada característica no universo de postagens consideradas.

A avaliação destas métricas serve para o entendimento acerca do comportamento geral dos usuários sobre determinado ambiente de conversa. A identificação dos picos é feita pela visualização da série temporal da quantidade de postagens.

O Twitter possibilita a utilização de métricas específicas que denotam os diferentes tipos de comportamento de seus usuários, dentre elas, ressalta-se a penetração. As características avaliadas são as seguintes:

- RT - *Tweets* que repassaram uma mensagem que já havia sido postada anteriormente por outro usuário;
- @ SEM RT - Direcionamento de mensagens para outra pessoa, que não foi um *retweet*;
- HTTP - *Tweets* que possuem informações contidas em sites da internet;
- HASHTAG (#) - Grupo de discussão sobre algum assunto específico.

5.3 Avaliando correlações

A avaliação de correlações é uma análise essencial para verificar as relações existentes entre indicadores diversos. Neste estudo buscou-se utilizar o coeficiente de correlação de Pearson para identificar a relação existente entre a intenção de voto e o índice de sentimento acerca dos candidatos no Twitter.

O´Connor et al. (2010) sugerem que a correlação entre índices de pesquisas e informações do Twitter devem ser avaliadas considerando a média móvel no vetor de informações oriundas da rede social. Tal procedimento é justificado mais pela maior variabilidade dos dados temporais oriundos do Twitter do que das informações vindas da pesquisa política. Contudo, neste artigo, realizou-se a correlação diretamente, uma vez que a quantidade de pontos existentes para a avaliação é muito reduzida devido a limitação da quantidade de pesquisas. Buscou-se elaborar um índice de sentimento no Twitter com uma ideia próxima ao de uma eleição, considerando como votos as postagens positivas sobre um candidato. Logo, a proporção de postagens positivas foi o primeiro indicador de interesse a ser testado. O cálculo foi feito considerando a performance das três principais candidaturas: PT, PSDB e PSB. Considerou-se, neste estudo, a candidatura de Marina Silva e Eduardo Campos, ambos da coligação liderada pelo partido PSB, como uma única candidatura. As intenções de votos das três candidaturas somam 100%, a mesma ideia foi aplicada quanto a proporção de

postagens. Outro ponto importante é que, para o cálculo das correlações, só foram consideradas as postagens que tiveram apenas um único candidato mencionado.

5.4 Análise semântica

Para a análise dos assuntos comentados pelos usuários sobre política, foi utilizada a análise de tópicos. O modelo Latent Dirichlet Allocation (LDA) é uma estrutura probabilística generalizada para a modelagem de matrizes esparsas de dados de contagem, tais como as matrizes utilizadas em *Text Mining* (*Bag of words*). A principal ideia por trás desse algoritmo é que as palavras de cada documento são geradas por uma mistura de temas (tópicos).

Segundo (Chein, 2016), o LDA representa os documentos (no caso, postagens) como uma mistura de tópicos, na qual, cada palavra é alocada a um tópico com uma probabilidade definida. Este modelo assume que, cada postagem, é produzida do seguinte modo: quando o autor (usuário) escreve um documento ele decide:

- Primeiro os tópicos a serem escritos;
- Depois a escolha das palavras para escrever sobre os tópicos (de acordo com uma distribuição multinomial);
- A quantidade de palavras *N* que o documento deve conter (distribuição de Poisson);
- A quantidade de mistura de tópicos que cada documento deve conter (distribuição de Dirichlet por meio dos *K* tópicos). Neste caso, um documento pode conter mais de um tópico diferente;
- A geração de cada palavra no documento.

O modelo de análise de tópicos LDA busca, recursivamente, obter a probabilidade de que um conjunto de tópicos tenha gerado os documentos da coleção. A estimação dos parâmetros que compõe o modelo é feita pelo método Collapsed Gibbs Sampling.

A análise da correlação entre tópicos foi feita conforme o seguinte processo: primeiramente, foi realizada a análise léxica. No segundo momento, foi feita a limpeza de *stopwords* (palavras sem valor semântico), para posterior execução do algoritmo de *stemming* (extração de radicais). Após esses passos, a matriz BOW foi calculada. Nesta matriz, cada termo considerado corresponde a uma coluna, e cada linha a um documento (*tweet*). A medida utilizada para avaliação foi o tf-idf (*term frequency inverse document frequency*).

As postagens relativas às ondas 9 e 14, realizadas imediatamente antes dos turnos da eleição, foram avaliadas por meio da Análise de Tópicos utilizando o modelo LDA.

A modelagem LDA possibilita a identificação das principais palavras relacionadas a cada um dos tópicos. Pela interpretação destas principais palavras pertencentes a cada tópico, é feita a interpretação de seu significado.

6 Coleta de dados

Para este artigo, foram consideradas todas as postagens do Twitter sobre os candidatos considerados nos dias nos quais foram realizadas as pesquisas. Foram realizadas, ao todo, 14 ondas: 10 no primeiro turno e 4 no segundo. A Tabela 1 mostra a relação de dias em que cada uma das ondas foi realizada.

Tabela 1 - Intervalos de tempo considerados

Turno	Onda	Data do campo
1º	1	18 a 21/07/2014
	2	03 a 06/08/2014
	3	23 a 25/08/2014
	4	31/08 a 02/09/2014
	5	13 a 15/09/2014
	6	20 a 22/09/2014
	7	27 a 29/09/2014
	8	29/09 a 01/10/2014
	9	02 a 04/10/2014
	*10	05/10/2014
2º	11	07 e 08/10/2014
	12	14/10/2014
	13	20 a 22/10/2014
	14	24 e 25/10/2014

Ao todo foram coletados 3.096.032 tweets sobre os candidatos Dilma, Aécio, Eduardo Campos e Marina nos intervalos mencionados. Todos eles foram classificados conforme o modelo de análise de sentimento desenvolvido. Contudo, para análise de correlação foram considerados 2.388.300, pois continham o nome de apenas um único candidato.

Quanto às pesquisas, os dados foram coletados seguindo o plano amostral desenvolvido pelo IBOPE Inteligência.

7 ANÁLISE DE DADOS

7.1 Acurácia do modelo de análise de sentimento

Na Tabela 2, pode ser vista a série histórica de acompanhamento mensal das métricas de ajuste obtidas para o modelo de análise de sentimento.

Tabela 2 - Qualidade do ajuste do modelo de análise de sentimento

Métrica	jul-14	ago-14	set-14	out-14
Recall	0,77	0,72	0,75	0,79
Precision	0,77	0,78	0,81	0,84
Accuracy	0,73	0,71	0,72	0,74
F-measure	0,77	0,75	0,78	0,82

A métrica *Recall* representa o percentual de verdadeiros positivos dentre o total de falsos negativos e verdadeiros positivos. A *Precision* é relativa a proporção de verdadeiros positivos dentre o total de falsos positivos e verdadeiros positivos. A métrica *Accuracy* representa o total de acertos dentre o total de casos possíveis. Por último, a *F-measure* é a média harmônica entre *Recall* e *Precision*. A Figura 3 aponta a matriz de confusão.

Matriz de confusão		Real	
		Sim	Não
Previsto	Sim	Verdadeiro Positivo	Falso Positivo
	Não	Falso Negativo	Verdadeiro Negativo

Figura 3 - Matriz de confusão

Comparando estes resultados com os obtidos por Araújo, Gonçalves e Benevenuto (2013), pode-se concluir que os níveis de assertividade do algoritmo de análise de sentimento elaborado são, inclusive, superiores em relação a um grande número de soluções disponíveis no mercado, para língua inglesa, pois possui uma quantidade de estudos desenvolvidos muito superior aos de língua portuguesa.

7.2 Evolução de postagens sobre os candidatos

Durante o período analisado, constatou-se um aumento na quantidade de postagens dentre a primeira onda e a última, sendo os picos ocorridos nas ondas 9 e 14, dias imediatamente anteriores aos turnos das eleições, conforme pode ser visto na Figura 4.

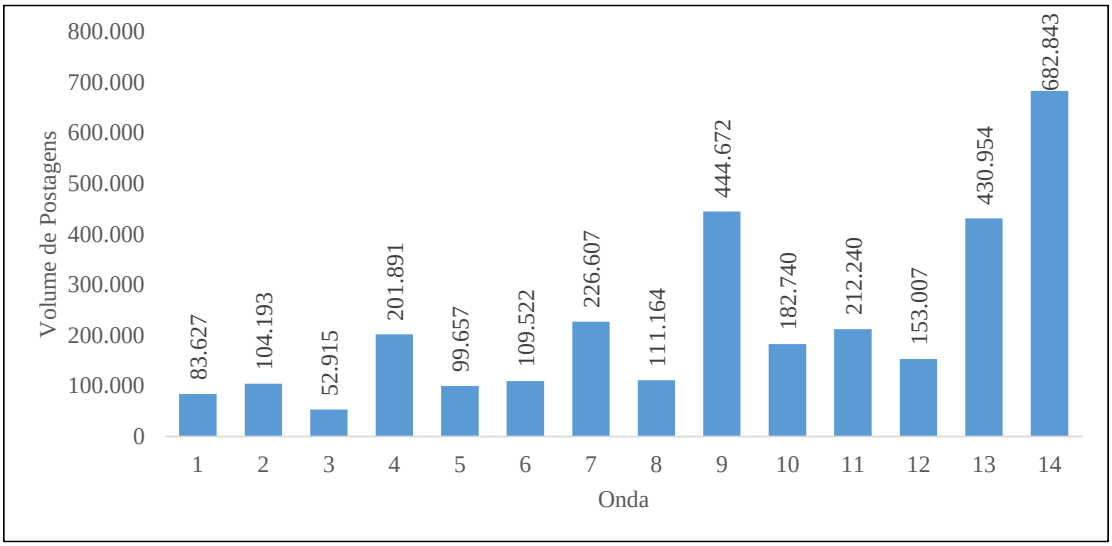

Figura 4 - Evolução na quantidade de postagens sobre os candidatos

Pela Figura 5, é possível observar uma tendência de aumento também na proporção de *retweets*, 62% na onda 1, chegando a 76% na onda 14. Esta alta proporção em toda a série histórica mostra um forte repasse de informações, o que configura uma característica do assunto política no Twitter: poucos geram informação e muitos repassam.

A proporção de *hashtags* (#) também apresenta aumentos significativos variando entre 15% e 48%, mostrando a evolução da popularização do tema eleitoral no Twitter.

Já a proporção de HTTP apresenta maior representatividade nas primeiras três semanas e depois decresce. Contudo, nunca apresenta valores inferiores a 26%, média de 43% entre a 4ª e a 14ª onda, valor que pode ser considerado alto, indicando a presença de postagens que direcionam a informações existentes em sites. Já a proporção de "@ sem RT", apresenta certa regularidade durante todo o período.

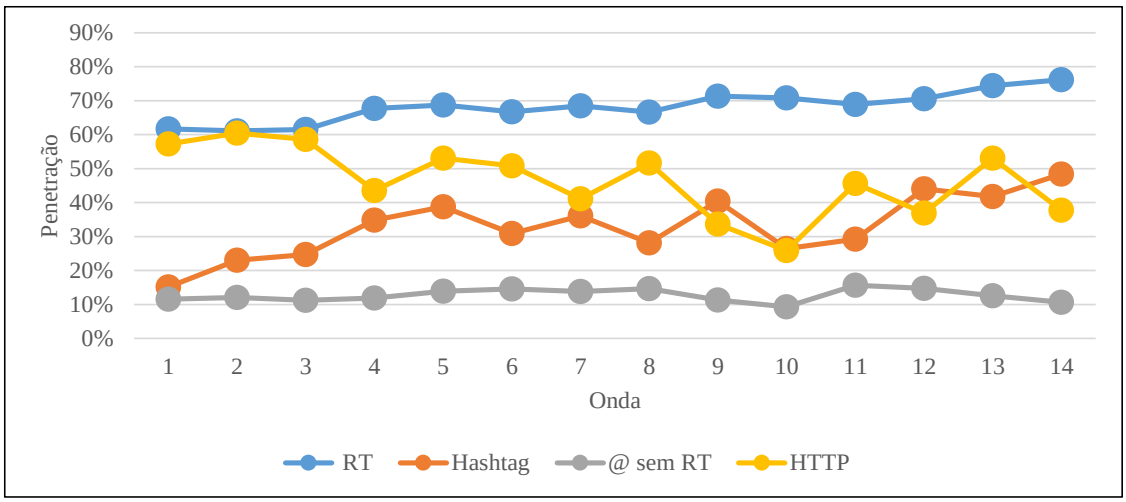

Figura 5 - Evolução de métricas do Twitter

7.3 Perfil dos usuários que interagem com a campanha pelo Twitter

Dentre os usuários que falaram sobre política no Twitter no período considerado no estudo, pode-se perceber que, 52% realizaram uma única postagem. Os usuários que fizeram mais de 50 postagens (9.420 usuários) foram responsáveis por 48% do total de postagens realizadas e 70% das impressões (Tabela 3).

Tabela 3 - Distribuição da qualidade postagens por usuário

Quant. de postagens	Usuários	%	Postagens	%	Impressões	%
1	243.127	52%	243.127	8%	260.616.867	3%
2	77.280	16%	154.560	5%	222.381.564	2%
3	37.898	8%	113.694	4%	170.847.033	2%
4	22.561	5%	90.244	3%	120.361.596	1%
5	14.835	3%	74.175	2%	95.079.530	1%
6 a 10	33.088	7%	248.606	8%	396.183.962	4%
11 a 20	19.271	4%	278.360	9%	622.388.884	6%
21 a 50	12.577	3%	393.974	13%	1.085.334.096	11%
51 ou mais	9.420	2%	1.499.292	48%	7.032.772.640	70%
Total	**470.057**	**100%**	**3.096.032**	**100%**	**10.005.966.172**	**100%**

Fazendo o *ranking* de usuários pelo volume de impressões (quantidade de postagens X quantidade de seguidores), conforme mostra a Tabela 4, percebe-se que o usuário dilmabr (oficial da candidatura petista) foi o que mais gerou impressões, tendo realizado 436 postagens no Twitter no período considerado. Usuários de agências de notícias aparecem até a quinta posição, tais como *G1*, *Jornal O Globo*, *Portal R7* e *Revista Veja*.

A primeira personalidade aparece na sexta posição, o apresentador Danilo Gentili, que realizou 43 postagens sobre os candidatos no período.

O usuário silva_marina (oficial da candidatura de Marina Silva) aparece na décima terceira posição. A falta de engajamento do candidato Aécio Neves no Twitter como criador de conteúdo pode ser apontado como um dos fatores de explicação da grande diferença entre suas proporções de postagens e a da candidata petista no início de sua campanha, mas com a proximidade das eleições seu nome ganha uma repercussão maior.

Vale ressaltar que a quantidade de seguidores considerados na Tabela 4 é referente ao período considerado na análise. Atualmente, os volumes são maiores.

Tabela 4 - Top de usuários por volume de impressões

Rank	UserID	NomeUsuario	Seguidores	Postagens	Impressões
1	89826	dilmabr	2.335.703	436	1.018.366.508
2	11435	g1	2.801.001	141	394.941.141
3	12091	JornalOGlobo	1.493.103	198	295.634.394
4	14448	portalR7	3.070.130	88	270.171.440
5	20450	VEJA	3.322.907	79	262.509.653
6	19511	DaniloGentili	5.917.726	43	254.462.218
7	11926	Estadao	1.244.307	150	186.646.050
8	64886	PastorMalafaia	741.810	220	163.198.200
9	14937	folha_com	1.322.629	122	161.360.738
10	147778	DaviSacer	398.124	397	158.055.228
11	11423	Val_Ce1	152.675	740	112.979.500
12	11720	TerraNoticiasBR	634.838	172	109.192.136
13	16836	silva_marina	838.921	127	106.542.967
14	21531	drangelocarbone	1.627.457	59	96.019.963
15	269254	rodrigovesgo	5.049.476	19	95.940.044
16	368839	lobaoeletrico	262.346	296	77.654.416
17	14274	UOLNoticias	454.990	169	76.893.310
18	13806	cartacapital	476.360	118	56.210.480
19	47350	felipeneto	2.732.792	17	46.457.464
20	10771	massavcs	144.700	319	46.159.300

7.4 Perfil do eleitor

Os eleitores brasileiros respondentes das pesquisas do IBOPE são, em sua maioria, mulheres 52%. Possuem maior concentração na faixa etária dos 25 a 34 anos, 25%. O ensino médio é o grau de instrução predominante, 39%. Já o ensino primário e fundamental, juntos, somam 42% do total de eleitores. As amostras das pesquisas realizadas possuem abrangência nacional com tamanho mínimo de 2.002 entrevistas, o que corresponde a um erro amostral estimado de 2 pontos percentuais.

Em relação a distribuição por região, verificam-se as diferenças existentes entre os eleitores brasileiros e as postagens sobre as três candidaturas no período.

Vale ressaltar que nem toda postagem captada contém a sua geolocalização, apenas 36% possui esta informação. Afinal, nem toda postagem é emitida por meio de dispositivos portáteis que possibilitem esta identificação.

Assumindo que a distribuição por região das postagens geolocalizadas seja igual às que não foram geolocalizadas, realizou-se o comparativo da Tabela 5 com o público da pesquisa.

Pode-se verificar que as diferenças não foram tão discrepantes, ou seja, os usuários do Twitter que postaram mensagens sobre os candidatos analisados estão dispostos de maneira próxima, na mesma ordem de grandeza, em relação a distribuição da quantidade de eleitores. A maior diferença, em termos de representatividade, está nas proporções das regiões Nordeste e Sudeste.

Tabela 5 - Distribuição de postagens e eleitores por região

Região	Pesquisa	Twitter	Total
Sudeste	43%	51%	50%
Nordeste	27%	20%	21%
Sul	15%	15%	15%
Centro-Oeste	8%	9%	9%
Norte	7%	5%	5%
Total	100%	100%	100%

7.5 Correlações históricas

Foram avaliadas as correlações entre as séries históricas de intenção de votos e proporção de postagens positivas para as três candidaturas e verificadas as correlações positivas em todos os casos. Contudo, observa-se o mesmo fenômeno quando se calcula o indicador pelas postagens com polaridade negativa, ou seja, quanto maior é a negatividade, maior a intenção de voto do candidato. Conclui-se então, que a correlação é alta, independentemente da polaridade da postagem. Analisando as séries históricas da representatividade da candidata Dilma nas postagens realizadas, conclui-se que sua proporção é praticamente a mesma, embasando a conclusão supracitada.

Vale lembrar que a representatividade mostrada na Figura 6 é a comparativa entre os candidatos. Por exemplo, na onda 1, a candidata Dilma teve 85,2% do total de postagens positivas, 84,2% do total postagens negativas, mesmo valor dentre as neutras. Percebe-se que valores próximos ocorrem em praticamente todas as ondas. Este mesmo fenômeno se repetiu nos demais candidatos.

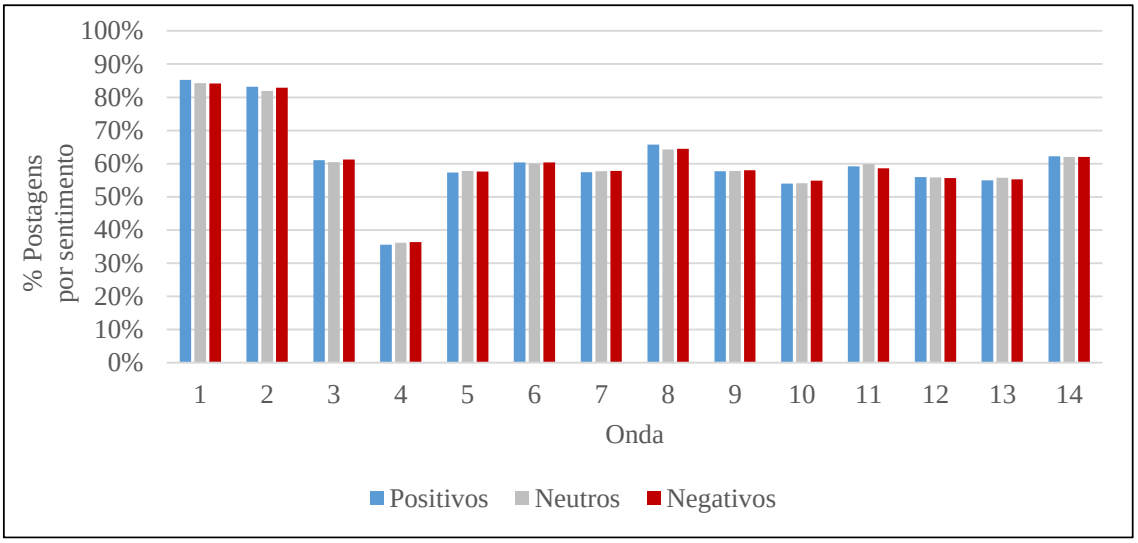

Figura 6 - Representatividade da candidata Dilma no total de postagens por sentimento

Com base no conhecimento adquirido, passa-se a avaliar as correlações com base no total de postagens realizadas, não apenas as positivas. Inicialmente, considerando as 10 ondas de pesquisas feitas no primeiro turno, a correlação do candidato Aécio se aproxima de zero, pois o mesmo apresentou altas taxas de intenção de voto e baixa proporção de postagens nas duas ondas iniciais, quando a campanha presidencial ainda estava "morna". Contudo, se considerar-se apenas as ondas que ocorreram após a morte de Eduardo Campos e do lançamento da candidatura de Marina Silva, verifica-se que a correlação aumentaria para 0,65. Avaliando-se o primeiro e o segundo turnos juntos, com as proporções relativas apenas à Dilma e ao Aécio, verificam-se correlações altas. Considerando-se todo o histórico, a correlação foi de 0,66 para ambos (Tabela 6).

Tabela 6 - Correlações de Pearson

Período	Aécio	Dilma	Eduardo/Marina
10 Turno - 10 Ondas	0,01	0,77	0,95
10 Turno e 20 Turno -14 Ondas	0,66	0,66	

Analisando a série histórica relativa ao primeiro turno, a candidata Dilma, atual presidente, apresentou alta proporção de postagens e intenção de votos nas primeiras duas semanas, possivelmente por ser o nome mais conhecido, sendo as proporções de postagens bem superiores às de intenção de voto. Já na pesquisa de Boca de Urna (onda 10) é possível ver que os indicadores ficam bem próximos (Figuras 7, 8 e 9).

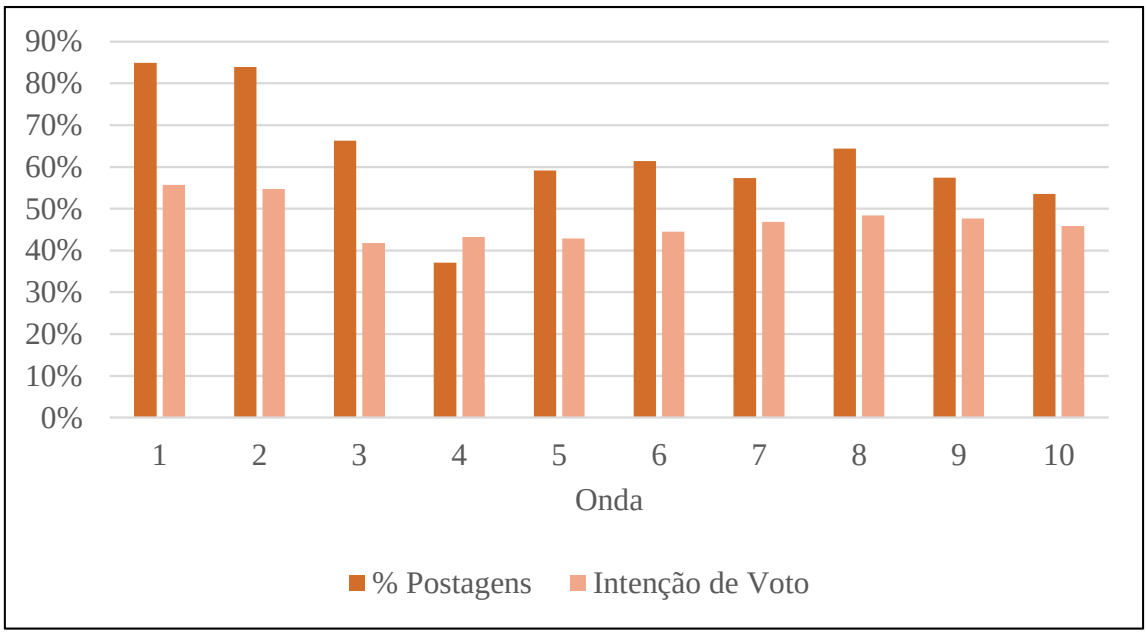

Figura 7 - Percentual de postagens e intenção de voto no primeiro turno da candidata Dilma

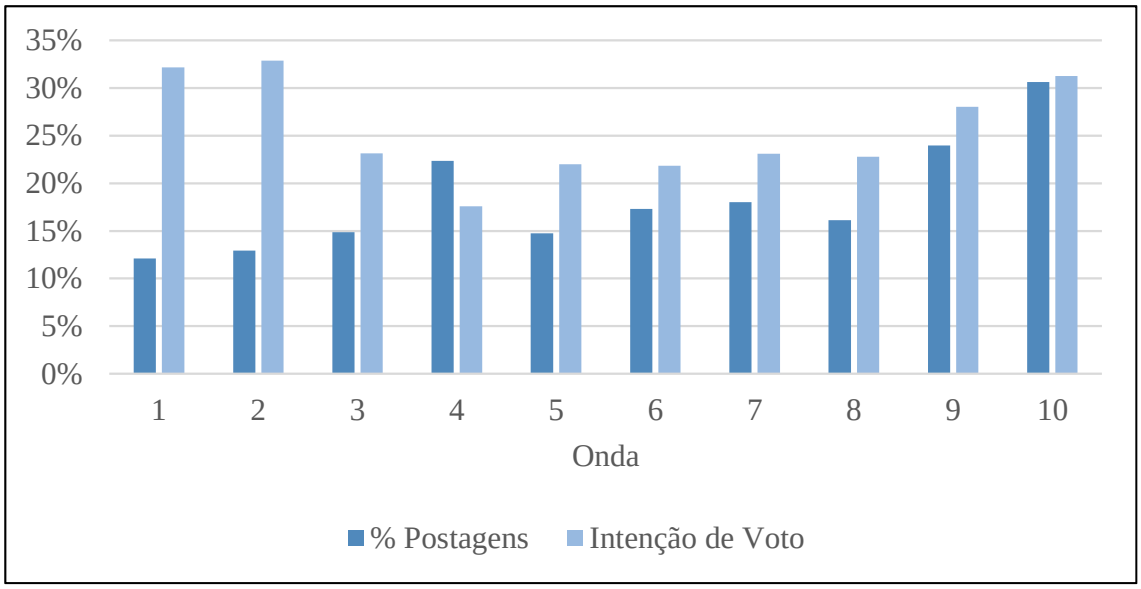

Figura 8 - Percentual de postagens e intenção de voto no primeiro turno do candidato Aécio Neves

Já Aécio Neves, apresentou uma participação muito inferior a Dilma no Twitter no primeiro turno das eleições. Principalmente nas primeiras duas semanas.

Na Figura 9 referente a candidatura de Eduardo Campos/Marina Silva para o primeiro turno, observa-se que as informações do Twitter possuem comportamento próximo ao observado nas pesquisas, tendo seu pico observado na onda 4, segunda pesquisa após o lançamento da candidatura de Marina, que decaiu nas semanas seguintes.

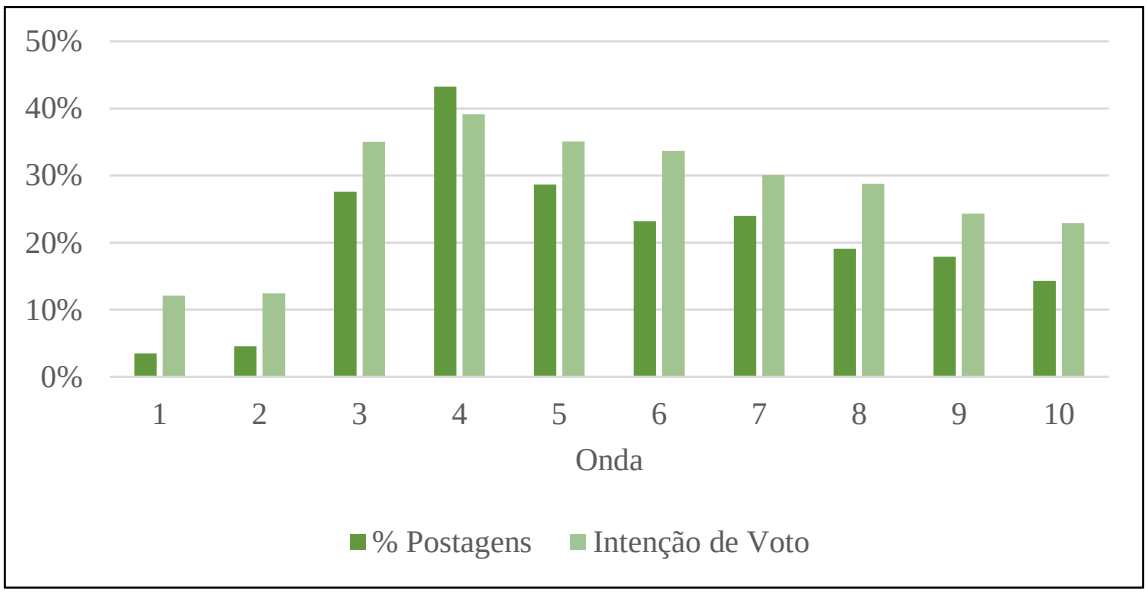

Figura 9 - Percentual de postagens e intenção de voto no primeiro turno dos candidatos Eduardo Campos/Marina Silva

Considerando-se as estatísticas das três candidaturas conjuntamente (10 x 3 = 30 pontos) no primeiro turno, obtém-se uma correlação de 0,92 entre intenção de voto e proporção de postagens no Twitter. Gerando um modelo de regressão linear simples, foi possível verificar que o aumento de 1 ponto percentual na representatividade do candidato em relação aos demais no Twitter gerou, em média, um aumento de 0,4783 pontos percentuais na intenção de voto do candidato durante o primeiro turno. A regressão se mostrou significativa, com p-valor interior a 0,001. Logicamente, este modelo se trata de uma aproximação, não sendo preciso para captar pequenas diferenças entre candidatos durante o processo eleitoral. Contudo, por meio dele comprova-se a forte relação existente entre as métricas durante a campanha.

Realizando a mesma avaliação, mas considerando apenas Aécio e Dilma (conjuntamente) para as 14 ondas, obtém-se resultados semelhantes. Um modelo com R^2 de 0,83 (Figuras 10 e 11).

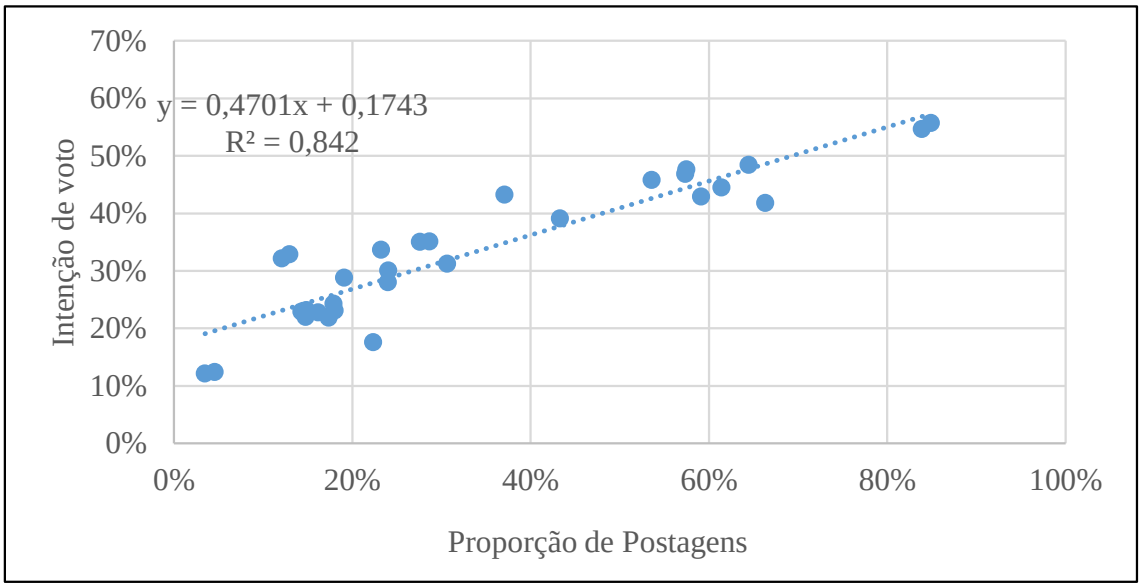

Figura 10 - Regressão Linear Simples - Primeiro Turno - Três candidaturas

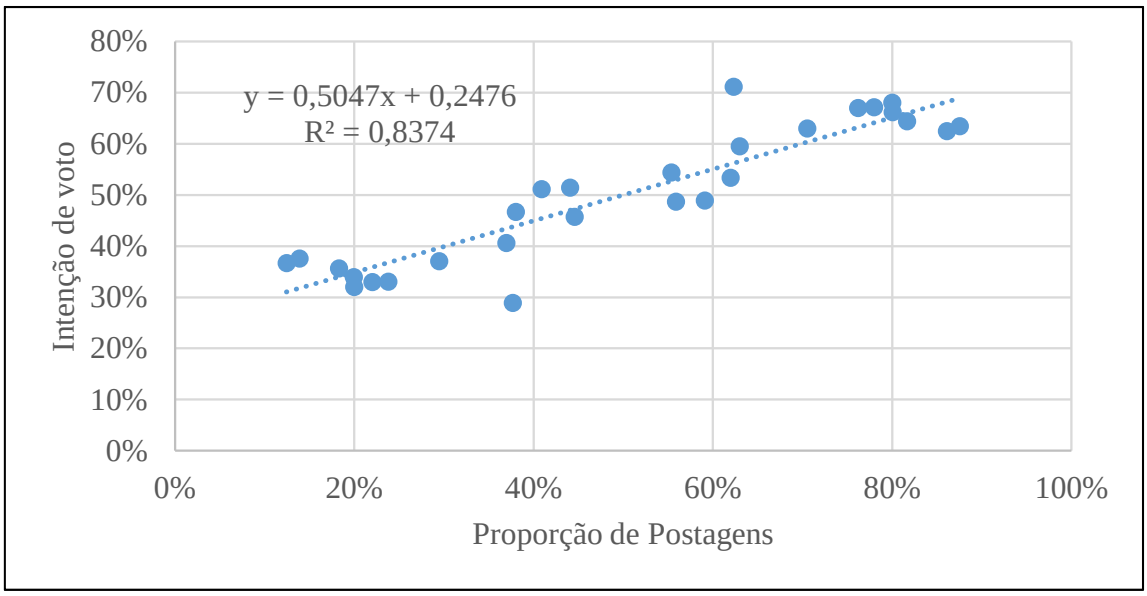

Figura 11 - Regressão Linear Simples – Primeiro e Segundo Turno - Duas candidaturas

A análise dos gráficos foi feita considerando apenas os votos de Dilma e Aécio (obtidos nas pesquisas) e postagens dos mesmos candidatos no Twitter. Pode-se perceber que a tendência é a mesma, Dilma inicia com uma proporção muito maior e esta proporção tende a convergir com a de Aécio no segundo turno. Este resultado é mais um indício da relação existente (Figura 12).

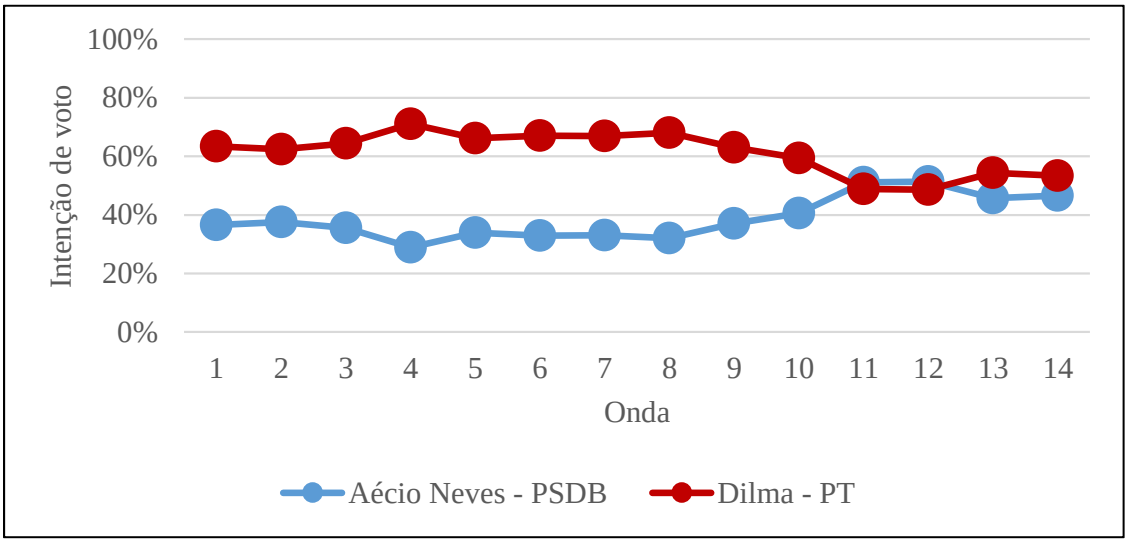

Figura 12 - Evolução Aécio e Dilma – Pesquisas

A maior diferença em relação a leitura ocorre na onda 4, quando a proporção de postagens de Aécio sobe no Twitter, mas decresce em relação a sua representatividade em relação a candidata petista nas pesquisas de intenção de voto (Figura 13).

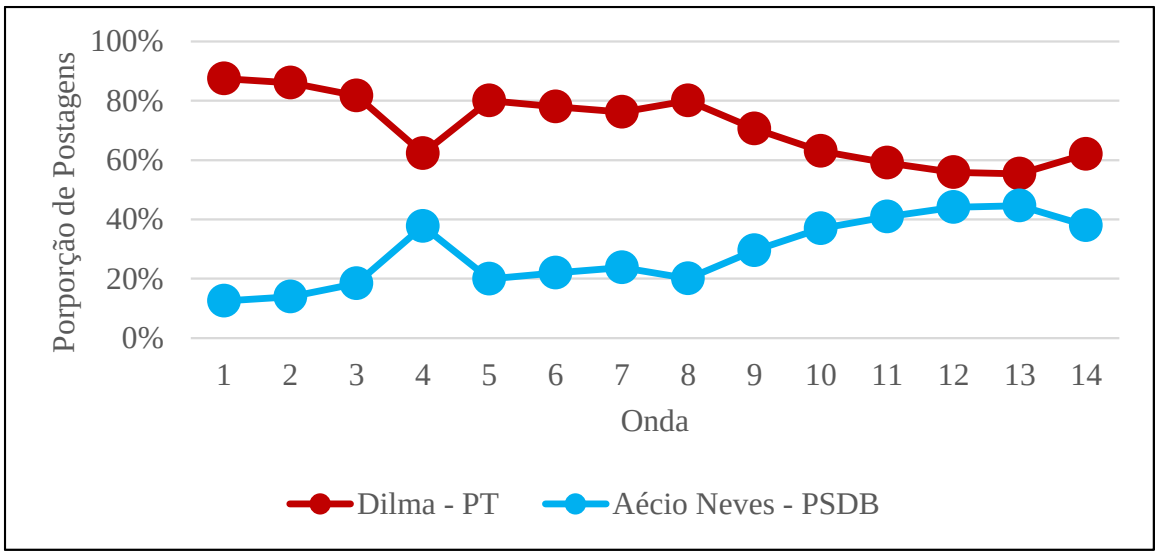

Figura 13 - Evolução Aécio e Dilma – Postagens

7.6 Tópicos relacionados

A análise de tópicos foi feita considerando uma amostra aleatória de 20 mil postagens para as ondas 9 e 14 (10 mil para cada uma), às vésperas dos turnos das eleições. Optou-se pela identificação de cinco tópicos em cada uma. A escolha do número de tópicos foi feita com base na interpretação dos resultados. Para cada tópico, selecionaram-se as 15 palavras mais pertinentes para sua identificação.

Avaliando-se a Figura 14, é possível verificar as palavras mais representativas dos tópicos da onda 9. O primeiro tópico contém, predominantemente, postagens de apoio ao candidato Aécio Neves, denunciando escândalos envolvendo o partido da candidata petista. Já no segundo tópico, estiveram mais presentes as razões para se optar pela candidata Marina Silva. No terceiro, Dilma aparece como figura central, sendo o controle da inflação a argumentação de destaque de sua defesa, aparecendo também fortes negativas à sua candidatura, tais como as postagens feitas pelo pastor da igreja Assembleia de Deus Silas Malafaia (#chegaderoubalheiraforadilma). No quarto tópico, aparecem postagens relacionadas a outros candidatos, tais como Luciana, Levy e Pastor Everaldo. O quinto tópico relaciona, principalmente, postagens especulativas sobre o resultado das eleições, mencionando também o debate da Rede Globo.

Onda 9: Tópicos				
1	**2**	**3**	**4**	**5**
aecioneves	**40**	**dilmabr**	**marina**	**marina**
dilmabr	**silvamarina**	**chegaderouba...**	silva	**turno**
45aecioconfir...	domingo	**presidenta**	**luciana**	**presidente**
corrupção	**votar**	vai	pra	neves
pt	conheça	novo	**levy**	aécio
sobre	**razões**	**marina**	neves	aecioneves
aécio	httptcoaz...	pastormalafaia	pergunta	ser
petrobras	neste	pra	candidato	silva
diz	vou	**pt**	**pastor**	**segundo**
correios	**brasilmarina40**	**inflação**	aécio	pode
minas	**marina40**	13	**everaldo**	**debate**
neves	dia	**controle**	corrupção	pois
educação	fazer	**dilma**	noite	**pesquisa**
fhc	dias	presidente	falar	**globo**
frase	econômica	**povo**	**vai**	qualquer

Figura 14 - Tópicos da onda 9

Já na onda 14, conforme mostra a Figura 15, o primeiro tópico esteve relacionado a postagens de usuários que apoiaram Aécio e os que apoiaram Dilma. Os defensores de Aécio acusavam Dilma de irregularidades relacionadas ao empréstimo do BNDES para o porto de Cuba. Já os aliados de Dilma, ressaltavam a construção de escolas ter sido superior no governo petista em relação ao governo do PSDB. No segundo tópico, ressaltaram postagens relacionadas ao pedido de direito de resposta da coligação de Dilma à *Revista Veja*. No terceiro tópico, aparecem postagens do usuário @OGloboPolítica avaliando se as falas dos candidatos no debate foram verdadeiras ou não. No quarto foi possível verificar postagens de repúdio à candidata petista e apoio à candidatura de Aécio Neves. Já no quinto tópico, o apoio do jogador Neymar à candidatura de Aécio Neves e a comparação de Aécio Neves a Fernando Henrique foram os destaques. Pode-se observar a pertinência dos tópicos considerados, por meio da interpretação de resultados. A análise LDA foi feita utilizando o programa R.

Onda 14: Tópicos				
1	**2**	**3**	**4**	**5**
governo	**veja**	vida	brasil	neves
anos	**tse**	vai	**mensalão**	presidência
quer	**revista**	tirar	**danilogentili**	candidato
aecioneves	eleitoral	pra	pastormalafaia	momento
brasil	lula	brasil	presidente	qualquer
vcs	jornaloglobo	oglobopolitica	é	ser
somostodos...	**pedido**	somostodos...	nunca	pode
psdb	**dilma**	eleição	pra	neymar
cuba	**resposta**	**pretonobranco**	**elessabiam...**	preso
escolas	**direito**	governo	mineiro	aécio
maior	nega	corrupção	corrupto	chamo
brasileiros	critica	eleitor	mudança	fernando
educação	terrorismo	checa	foradilma	henrique
esconder	fundadora	debate	aecio45pelo...	aecio
porto	justiça	debatenaglobo	pq	eaecio45...

Figura 15 - Tópicos da onda 14

8 CONCLUSÕES

Verifica-se a pertinência das informações oriundas do Twitter como importante fonte complementar de análise às pesquisas eleitorais realizadas, apresentando como principais vantagens a menor granularidade de tempo e possibilidade de interpretação de resultados quase instantaneamente. Contudo, a proporção de postagens não pode e não deve ser utilizada para estimar a proporção de votantes de determinado candidato. Para isso, as pesquisas apresentam resultados muito mais coerentes, dado que nelas o respondente é exposto a perguntas objetivas no qual escolhe o candidato em que mais possui afinidade considerando determinado cenário estabelecido. Tal procedimento inexiste nas redes sociais, nas quais o usuário não possui limites para a exposição de suas ideias e opiniões.

Outro aspecto relevante é que a distribuição de postagens e eleitores nas regiões brasileiras foi próxima. Este fato, de maneira alguma, pode indicar que exista uma semelhança em relação aos demais aspectos sociodemográficos. Contudo, constitui um indício interessante para que haja uma investigação futura mais profunda.

Em relação ao perfil dos usuários que realizaram postagens sobre os candidatos no período considerado, pode-se dizer que, cada um possui, em média, 1.382 seguidores, sendo esta distribuição bastante assimétrica, já que a mediana é de 183 seguidores. Os meios de comunicação se mostram importantes disseminadores das informações políticas, uma vez que estiveram dentre os usuários mais retuitados.

Quanto a análise de sentimento, verifica-se que o modelo de sentimento desenvolvido consegue captar de forma coerente a polaridade das postagens. Contudo, a informação gerada por ele não possui relação com as variações ocorridas nas intenções de voto captadas nas pesquisas IBOPE, como diria a música do Charlie Brown Jr.: "Falem bem, falem mal, mas falem de mim".

Em relação a análise de tópicos (modelo LDA), percebe-se adequação aos dados, sendo capaz de identificar tópicos pertinentes capazes de oferecer uma rápida interpretação das informações postadas. Soluções baseadas neste tipo de modelagem poderão fornecer uma avaliação mais rápida das notícias políticas nas próximas eleições.

9 LIMITAÇÕES E SUGESTÕES DE NOVAS PESQUISAS

Há de se considerar o fato das correlações serem feitas considerando-se um número muito pequeno de pontos, pela limitação existente no Brasil da quantidade de pesquisas eleitorais realizadas e divulgadas. Por este motivo, foram realizadas avaliações conjuntas dos candidatos, afim de aumentar a robustez das conclusões apresentadas.

Recomenda-se fortemente a realização de estudos como este em eleições futuras, com o intuito de verificar se as conclusões obtidas se manterão. A expectativa é de que as correlações se tornem cada vez maiores, devido ao aumento de acesso da população brasileira à internet e, consequentemente, à rede social Twitter. Contudo, trata-se de uma hipótese que deve ser verificada.

A utilização de informações de intenção de voto oriundas de outros institutos de pesquisa também pode ser uma opção promissora para análises futuras.

REFERÊNCIAS

Araujo, M., Gonçalves, P., & Benevenuto, F. (2013). *Métodos para análise de sentimentos no Twitter*. In Proceedings of the Simpósio Brasileiro de Sistemas Multimídia e Web (Web media).

Barion, E. C. N., & Lago, D. (2008). Mineração de textos. *Revista de Ciências Exatas e Tecnologia*.

Bourdieu, P. 2000 [1973]. La opinión pública no existe. *Cuestiones de Sociología*, 220-232. Madrid: Istmo.

Boyte, H. C. (1995). Public opinion as public judgement. In T. L. Glasser, & C. T. Salmon, (Eds.). *Public Opinion and the Communication of Consent*. Nueva York: The Guilford Press, 417-436.

Breiman, L. (2001). *Random Forests*. *Machine Learning*. *45*(1), 5-32. doi*:* 10.1023/A:1010933404324

Cavallari, M. (2016). *Congresso em Foco*. Entrevista. Recuperado de http://congressoemfoco.uol.com.br/noticias/marcia-cavallari-%E2%80%9Cpesquisa-nao-e-infalivel%E2%80%9D/

Chein, E. (2016). *Introduction to latent Dirichlet allocation*. Retrieved from http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/

Converse, J. M. (1987). *Survey research in the United States: Roots and emergence, 1890-1960*. University of California Press, Berkeley, California.

Gosh, R. A. (2016). *Social media for giant instant opinion polls: Twitter political index.* Retrieved from http://sentimentsymposium.com/SS2012w/presentations/SAS12w-RishabGhosh.pdf

Gramacho, W. G. (2015). Surveys pré-eleitorais nas eleições brasileiras de 2014: Erros, acertos e polêmicas. *REB - Revista de Estudios Brasileños*, Primer semestre, *2*(2), 115-113, Madrid: Universia.

Habermas J. (1998). *Facticidad y validez: Sobre el derecho y el estado democrático de derecho en términos de teoría del discurso*. Madrid: Editorial Trotta.

IBOPE Inteligência. (2016). *Avaliar a relação existente entre o índice de sentimento de candidatos no Twitter, durante a campanha eleitoral para presidência de 2014 e a intenção de voto dos principais candidatos*. Pesquisa. Recuperado de http://www.eleicoes.ibopeinteligencia.com.br

Lane, R. E., & Sears D. O. (1964). *Public opinion*. Englewood Cliffs: Prentice Hall, 13.

Lunden, I. (2012). *Analyst: Twitter passed 500m users*. In June 2012, 140M of them in US; Jakarta 'Biggest Tweeting' City. Retrieved from https://techcrunch.com/2012/07/30/analyst-twitter-passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city/

O'Connor B., Balasubramanyan, R., Routledge B. R., & Smith, N. A. (2010). *From tweets to polls: Linking text sentiment to public opinion time series.* Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.

Price, V. (1992). *Communication concepts 4: Public opinion*. Newbury Park: Sage Publications.

Ribeiro, R. O. A., Tavares, T. G. B., & Cohen, D. O. (2014). Análise de usuários que conversam sobre cerveja no Twitter. *PMKT - Revista Brasileira de Pesquisas de Marketing, Opinião e Mídia*, *14*, 174-195.

Salton, G., & Mcgill, M. J. (1983). *Introduction to modern information retrieval*. Computer Science Series, USA: McGraw-Hill.

SamPedro, V. (2000). *Opinión pública y democracia: Medios, sondeos y urnas*. Madrid: Istmo.

Sartori, G. (2002). *Elementos de teoría política*. Madrid: Alianza Editorial.

Sivic, J. (2009). *Efficient visual search of videos cast as text retrieval*. Transactions on pattern analysis and machine intelligence, *31*(4), IEEE.

Speier, H. (1950). Historical development of public opinion. *American Journal of Sociology*, *55*, 376-388.

Strachan, D. (2009). *Twitter: How to set up your account.* Retrieved from http://www.telegraph.co.uk/travel/4698589/Twitter-how-to-set-up-your-account.html

Tribunal Superior Eleitoral. (2016). Número de candidatos para o cargo de Presidente da República em 2014. Recuperado de http://www.tse.jus.br/

Worcester, R. (1997). Public opinion and the environment. In M. Jacobs (Comp.). *Greening the Millennium? The new politics of the environment*. Oxford: Blackwell Publishers.