Acompanhamento de Campanha Eleitoral pelo Twitter

Marden B. Pasinato
COPPE/UFRJ
Rio de Janeiro, RJ, Brasil
marden@cos.ufrj.br

Carlos E. Mello
DCC/UFRRJ
Nova Iguaçu, RJ, Brasil
carlos.mello@ufrrj.br

Geraldo Zimbrão
COPPE/UFRJ
Rio de Janeiro, RJ, Brasil
zimbrao@cos.ufrj.br

***Abstract*—O processo eleitoral vem tomando proporções cada vez maiores, de maneira que a ciência do marketing é, hoje, uma das principais preocupações das equipes responsáveis pelas campanhas eleitorais. O crescente uso das redes sociais criou um novo canal entre candidatos e eleitores. Contudo, é possível afirmar que este canal tem sido pouco explorado. É um hábito para muitos usuários divulgar nas redes sociais suas opiniões a respeito dos candidatos e partidos políticos. Se as equipes de marketing eleitoral tivessem acesso direto a tais informações, poderiam utilizá-las como "termômetro" em suas campanhas. Entretanto, filtrar as opiniões dos usuários não é uma tarefa trivial devido à enorme quantidade de informações fúteis sendo compartilhada nas redes. Portanto, este trabalho propõe-se a extrair do Twitter informações relevantes, i.e., as opiniões dos usuários a respeito de um determinado candidato. Como estudo de caso, tweets referentes a um dos candidatos à prefeitura da cidade do Rio de Janeiro (RJ) foram coletados, durante a campanha eleitoral de 2012, e um filtro foi aplicado para diferenciar os tweets que transmitem a opinião dos usuários (relevantes) dos demais (irrelevantes). Técnicas de Aprendizado de Máquina foram empregadas com o intuito de efetuar tal filtragem, entre as quais estão Naive Bayes, Árvore de Decisão e Máxima Entropia. Em nossos resultados, onde atingimos acurácia média de até 75%, tiramos importantes conclusões a respeito dos fatores que interferem no desempenho de tais técnicas.**

***Keywords*—Marketing Eleitoral, Aprendizado de Máquina, Processamento de Linguagem Natural, Mineração de Redes Sociais.**

I. Introdução

Em tempos de eleições é de vital importância que as equipes de marketing eleitoral tenham ciência da opinião pública a respeito dos candidatos. Uma vez que a equipe saiba de antemão a opinião do eleitorado, estratégias de marketing poderão ser planejadas e implementadas com maior exatidão. Por exemplo, ao descobrir que determinada notícia, a respeito de um candidato ou partido, causou descontentamento entre os eleitores, a equipe poderá planejar estratégias de modo a amenizar seus efeitos negativos no público.

Com o advento das redes sociais, a Internet tornou-se o grande depósito de opiniões da sociedade, pois cada usuário conectado à rede, além de consumidor, tornou-se um produtor de conteúdo. Captar a opinião desses usuários a respeito dos candidatos e partidos pode ser um diferencial em uma eleição. Todavia, extrair tais informações de redes sociais não é uma tarefa trivial, visto que há inúmeros assuntos sendo debatidos, a maioria completamente divorciada da campanha. Logo, as equipes necessitam de uma ferramenta que seja capaz de filtrar as informações relevantes das irrelevantes.

Por informações relevantes, nos referimos às opiniões pessoais dos usuários a respeito de um candidato, sejam elas favoráveis ou desfavoráveis. Neste trabalho, realizou-se um estudo de caso tendo o Twitter como fonte de informações relevantes para o marketing eleitoral. Tweets contendo o nome de um dos candidatos à prefeitura da cidade do Rio de Janeiro (RJ) foram coletados, durante a campanha eleitoral de 2012, através da API do Twitter, e técnicas de Aprendizado de Máquina foram empregadas para filtrar os relevantes dos irrelevantes.

Filtrar tweets contendo a opinião dos usuários durante o curso de uma eleição do porte da do Rio de Janeiro é uma tarefa árdua. São milhões de eleitores, muitos deles provavelmente são usuários ativos do Twitter, divulgando suas opiniões, a respeito de diversos temas, diariamente. Em meio aos milhares de tweets é preciso distinguir entre os que transmitem algum tipo de opinião dos que enunciam simples fatos ou notícias. Dentre os que transmitem alguma opinião, é preciso ainda diferenciar aqueles que se direcionam ao candidato alvo daqueles que emitem opiniões a respeito de outros temas ou outras personalidades. É possível afirmar, portanto, que tal tarefa é praticamente impossível de ser realizada manualmente, dada a enorme quantidade de tweets.

Tendo isto em mente, optamos por uma abordagem computacional que encara este problema como um problema de classificação supervisionada, onde os tweets podem pertencer a duas classes: *relevantes* e *irrelevantes*. Desta maneira, foi possível comparar o desempenho de diversos classificadores em conjunto com vários métodos de representação vetorial de texto. Dentre os classificadores utilizados estão: Naive Bayes, Árvore de Decisão e Máxima Entropia. Os métodos de representação utilizados foram: Booleano, TF, TF-IDF e SVD.

Em nossos experimentos, atingimos índices de acurácia considerados satisfatórios dado o carácter preliminar deste trabalho. Mais importante do que encontrar um classificador e método que resultem no melhor desempenho, mostramos que este problema pode ser encarado como um problema típico de classificação de texto. Além disso, analisamos como as características do problema afetam o desempenho das técnicas empregadas.

Este trabalho está dividido em quatro seções, das quais esta é a primeira. Na seção II, comentaremos os trabalhos

dentro da literatura que envolvem mineração de dados em redes sociais com a finalidade de se obter conclusões sobre eleições. Na seção III, descreveremos como nossa base de tweets foi adquirida e rotulada em classes, isto é, como montamos nosso *corpus*. Na seção IV, apresentaremos como nossos experimentos foram conduzidos, bem como os métodos de representação e os classificadores utilizados. Na seção V, divulgamos os resultados do experimento e analisamos os fatores que interferem no desempenho dos classificadores. Por fim, na seção VI, concluímos apontando possíveis trabalhos futuros.

II. Trabalhos Relacionados

Muitos trabalhos utilizam o Twitter para realizar *Análise de Sentimento* [1]–[4]. Neste tipo de problema, deseja-se descobrir, de forma automática, se o tweet contém um sentimento positivo ou negativo, da parte do autor, a respeito de algo ou alguém. Normalmente, tais problemas são abordados como problemas de classificação supervisionada, onde um tweet pode pertencer a três possíveis classes: *sentimento positivo*, *sentimento negativo* ou *sem sentimento*.

Pesquisas nesta linha ganharam grande notoriedade devido a sua aplicabilidade para fins de marketing. Geralmente, procura-se descobrir se um tweet possui sentimento positivo ou negativo em relação a uma marca, personalidade ou produto específico. Contudo, a análise de sentimento pode ser facilmente aplicada para fins eleitorais. Em [5], os autores procuram aplicar a análise de sentimento, em tempo real, durante a campanha eleitoral de 2012 para a presidência dos EUA.

Tanto [6] quanto [7] analisaram os dados do Twitter procurando prever os resultado das eleições presidenciais da Alemanha, em 2009. O primeiro trabalho afirma que o número de tweets mencionando um determinado partido já é um bom indicador de quantos votos este partido terá nas eleições. No entanto, o segundo trabalho, fazendo uso de outras métricas, a maioria baseada na quantidade de *hashtags* mencionando um candidato ou partido, conclui que há uma fraca correspondência entre os resultados traçados a partir do Twitter e os resultados reais da eleição.

O trabalho mais completo da área é [8]. Nele, o autor analisa diversos trabalhos que se propuseram a prever os resultados de eleições com base nos dados do Twitter. Tais trabalhos foram comparados considerando a janela de tempo em que os tweets foram adquiridos; o método utilizado para filtrar os tweets relevantes dos irrelevantes; a maneira como a previsão dos resultados é realizada; e a métrica usada para avaliar quão boa foi a previsão. Por fim, [8] conclui que, apesar do entusiasmo presente na área, as previsões feitas com base nos tweets são muito frágeis e duvidosas. Um dos principais motivos para isto é o fato do Twitter ser uma fonte de dados muito enviesada para a opinião do público jovem, uma vez que este público superabunda as redes sociais em geral.

Sabendo da dificuldade de se prever o resultado das eleições através de uma fonte não muito confiável, este trabalho busca, ao invés disso, tentar extrair as opiniões dos usuários do Twitter e apresentá-las às equipes de marketing para que estas tomem as devidas providências a respeito do rumo da campanha. Ou seja, propomos uma ferramenta de apoio às eleições e não uma ferramenta de análise das mesmas.

Tabela I. Exemplo de tweets rotulados

Critério	Tweet	Classe
I	*(canditado)*, do *(partido)*, participa hoje de sabatina	Irrelevante
II	já perceberam que o *(pessoa)* faz a imitação do *(candidato)*? tenho raiva do *(pessoa)*	Irrelevante
III	quero parabenizar o *(candidato)* pelo debate e perguntá-lo quais seriam suas primeiras medidas se eleito.	Relevante

III. Base de Dados

Construímos um *corpus* contendo tweets relacionados a um determinado candidato à prefeitura da cidade do Rio de Janeiro (RJ), durante as eleições de 2012. Para tal, fizemos uso da API do Twitter onde foi possível coletar tweets contendo o nome do candidato em seu texto. Os tweets coletados passaram por um processo de eliminação de tweets repetidos e eliminação de retweets. Após este processo, foi formada uma base contendo mais de 30.000 tweets, onde cada tweet menciona pelo menos uma vez o nome do candidato escolhido como foco deste trabalho. Deste total de 30.000 tweets, 1.000 deles foram rotulados manualmente por alunos de pós-graduação da UFRJ entre as classes: *relevante* e *irrelevante*. Alguns critérios foram definidos para determinar se um tweet é *relevante* ou *irrelevante*, conforme descrito abaixo:

- Se o tweet se refere a alguma outra pessoa que tem o mesmo nome, ou partes do nome, do candidato; então ele é rotulado como *irrelevante*.
- Se o tweet é um texto objetivo, ou seja, apenas uma notícia ou um fato; então ele é rotulado como *irrelevante*.
- Se a opinião expressa no tweet não se refere ao candidato, mas sim a alguma outra pessoa, ocorrência ou coisa; então o tweet é rotulado como *irrelevante*.
- Se o tweet expressa uma opinião sobre o candidato ou sobre alguma decisão ou ação do candidato; então o tweet é rotulado como *relevante*.

A Tabela I apresenta alguns exemplos de possíveis tweets e como os mesmos seriam rotulados. Para facilitar o trabalho de rotulação, uma interface gráfica, apresentada na Figura 1, foi disponibilizada na Web. A interface permite ao rotulador selecionar três opções: a primeira é rotular o tweet como *relevante*; a segunda é rotular o tweet como *irrelevante*; e a terceira é utilizada caso o rotulador esteja em dúvida.

IV. Experimentos

Para um humano é relativamente simples ler um texto e entender se o mesmo conduz algum tipo de opinião, da parte do autor, sobre um determinado objeto. Entretanto, para um computador, identificar uma opinião em um texto é uma tarefa díficil. Uma possível abordagem é introduzir um conjunto de treinamento, onde há textos previamente rotulados, a fim de que a máquina possa identificar os padrões que sinalizam uma opinião.

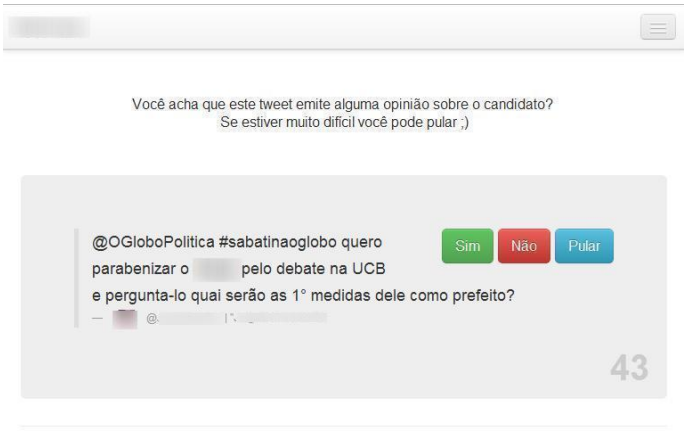

Figura 1. Interface para classificação manual dos tweets

Conhecida como classificação supervisionada, esta abordagem requer que os textos sejam mapeados em um espaço vetorial para que, a partir deste, seja possível identificar as regiões que caracterizam a presença ou ausência de uma opinião. O processo de classificação depende, portanto, do processo de criação do espaço vetorial [9].

A. Representação Vetorial

Há diversas formas de se representar um texto em um espaço vetorial. Contudo, existem quatro métodos de criação de espaço vetorial que são os mais comuns dentro da literatura de processamento de linguagem natural [10], [11].

1) Método Booleano
2) Método de Frequência de Termos (TF)
3) Método Frequência de Termos com Frequência Inversa de Documentos (TF-IDF)
4) Método de Extração de Variáveis Latentes (SVD)

Primeiramente, realiza-se um pré-processamento no corpus reduzindo o número de palavras que possuem fundamentalmente o mesmo significado. Por exemplo, substitui-se palavras acentuadas por suas versões sem acentos; remove-se a terminação de verbos e substantivos flexionados, deixando-se apenas os radicais; substitui-se letras maiúsculas por minúsculas; e, por fim, descarta-se palavras muito frequentes como preposições, artigos, verbos auxiliares, entre outras, conhecidas como *stopwords*.

As palavras e radicais remanescentes, também chamados de *tokens*, constituem o que se chama de *Bag-of-Words*, ou seja, o conjunto mínimo de elementos do corpus que transmitem algum significado. Calcula-se então quais tokens são os mais frequentes para que estes sejam usados como dimensões do espaço vetorial.

O método Booleano cria um vetor para cada tweet onde, em cada dimensão, verifica-se se o tweet contém ou não o token que esta dimensão representa. O método de TF age semelhantemente, porém calculando a frequência do token no tweet ao invés de simplesmente verificar sua presença. Por fim, o método TF-IDF calcula a frequência do token no tweet e multiplica a mesma por um fator penalizador, que leva em consideração a frequência do token dentro de todo o corpus.

O método de extração de variáveis latentes funciona de maneira distinta dos demais. Primeiramente, monta-se uma matriz A, onde as linhas representam os tweets e as colunas representam os tokens mais frequentes. Cada elemento a_{ij} da matriz representa o número de vezes que o token j aparece no tweet i. Aplica-se então a decomposição *Singular Value Decomposition* (SVD) [12], que retorna duas matrizes U e V.

As matrizes U e V representam, em k dimensões, os tweets e os tokens, respectivamente. Ou seja, se A for $n \times m$, havendo n tweets e m tokens, U será $n \times k$, enquanto que V será $m \times k$. O parâmetro k é arbitrário, sendo um dos parâmetros de entrada do algoritmo SVD. Ele representa o número de variáveis latentes nas quais os tweets e tokens serão mapeados.

Tais variáveis são consideradas latentes, pois não há como se chegar a um consenso sobre o que elas efetivamente representam. No caso dos métodos apresentados anteriormente (Booleano, TF e TF-IDF), cada dimensão do espaço vetorial representa um token específico. Entretanto, no caso das variáveis latentes, não se pode afirmar que elas, de fato, representam algo, pode-se apenas conjecturar sobre suas possíveis representações. Por exemplo, dado $k = 3$, podemos supor que as três variáveis representam o conteúdo jornalístico, o conteúdo emotivo e o conteúdo ficcional do tweet, embora outras interpretações sejam cabíveis.

Utilizar variáveis observáveis para representar o texto pode ser uma tarefa complicada. Geralmente, é necessário um grande número de variáveis observáveis a fim de se obter uma boa representação vetorial dos textos. Além disto, o cálculo de tais variáveis pode exigir um esforço computacional considerável, dependendo do tamanho do corpus (e.g., TF-IDF). Quando opta-se pela representação em variáveis latentes, ganha-se em simplicidade e flexibilidade, uma vez que o número de variáveis utilizadas é facilmente configurável e o método para extraí-las possui aplicação simples e direta.

De posse da representação dos tweets como pontos em um espaço vetorial, é possível agora trabalhar com técnicas de Aprendizado de Máquina voltadas para a tarefa de classificação. Realizamos, neste estudo de caso, uma comparação entre os classificadores Naive Bayes, Árvore de Decisão e Máxima Entropia.

B. Naive Bayes

Optamos por utilizar o Naive Bayes por questão de simplicidade [13]. É talvez a técnica de classificação mais simples dentro da literatura, porém está longe de ser uma técnica descreditada. Obtendo bons resultados com este classificador, teremos fortes indícios de que a abordagem clássica de Aprendizado de Máquina é uma abordagem válida para atacar o problema da filtragem de tweets.

O classificador Naive Bayes se baseia no conhecido teorema de *Bayes* [14]. Este teorema nos permite calcular a probabilidade condicional de um tweet pertencer a uma determinada classe. Calculando as probabilidades condicionais do tweet pertencer a cada uma de todas as possíveis classes, é razoável assumir que o mesmo pertencerá a classe cuja probabilidade condicional for a mais elevada.

C. Árvore de Decisão

A Árvore de Decisão também é considerada um classificador simples, entretanto ela é extensamente utilizada, em diversos tipos de aplicações, por retornar um resultado que pode ser interpretado como sendo a regra utilizada pelo classificador durante o processo de classificação [15]. Muitas vezes não queremos apenas classificar os dados em classes, e sim saber o porquê do algoritmo ter optado por uma classe em detrimento da outra. Na maioria dos classificadores, a tomada de decisão não é visível para o usuário, apenas o resultado final, o que nem sempre é útil.

O resultado fornecido pela Árvore de Decisão é uma árvore onde as folhas representam as classes e os nós intermediários representam as regras aplicadas a uma das variáveis envolvidas no problema. Podemos interpretar este resultado como sendo o processo de tomada de decisão que o algoritmo acatou, separando as instâncias pelas regras presentes nos nós intermediários. As instâncias que satisfazem a agregação das regras encontradas durante o percurso, da raiz à folha, são rotuladas como pertencendo à classe da folha em questão. Isto fornece ao usuário um melhor entendimento sobre a natureza de seus dados e lhe permite tomar decisões mais embasadas, ao invés de simplesmente confiar em uma "caixa preta".

D. Máxima Entropia

O classificador de máxima entropia procura definir, da maneira mais uniforme possível, as probabilidades de um elemento pertencer a cada classe, respeitando as restrições contidas no conjunto de treinamento. Por exemplo, numa tarefa de tradução automática de texto, uma palavra x de uma determinada língua pode ser traduzida, dependendo do contexto, em três possíveis palavras (y, z, w) de uma outra língua.

A princípio, como não há restrições, o modelo de máxima entropia definiria probabilidades iguais de x ser traduzida para cada uma das três possíveis palavras na outra língua. Entretanto, imagine que, no corpus de treinamento, 50% das vezes que em x é traduzida, ela é traduzida como sendo y. De posse desta informação, não faz mais sentido atribuir a mesma probabilidade a todas as classes, deve-se então recalcular as probabilidades incorporando no sistema de equações esta restrição.

Na prática, o número de restrições é enorme e resolver o modelo de máxima entropia consiste em resolver um problema de otimização que busca atribuir probabilidades, de forma mais uniforme possível, dado um conjunto de restrições. Conforme mostrado em [16], a extensão do conjunto de restrições impede uma resolução analítica do problema, portanto as técnicas utilizadas para se resolver o modelo de máxima entropia são técnicas que buscam uma solução aproximada, como gradiente descendente, entre outros.

E. Configurações

Percebemos que há vários possíveis espaços vetoriais para se representar os tweets considerando os possíveis métodos de criação e o número de dimensões a se considerar. O desempenho dos classificadores pode variar dependendo do espaço vetorial utilizado, portanto devemos testar qual a *configuração*, i.e., escolha do método e do número de dimensões, que melhor favorece os classificadores.

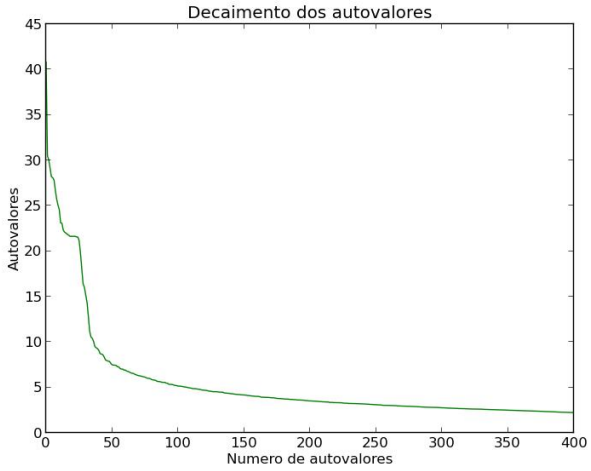

Figura 2. Decaimento dos autovalores

No caso do método de representação SVD, um dos parâmetros a ser definido é o número de variáveis latentes k que determina o número de dimensões do espaço vetorial. Utilizar muitas variáveis latentes agrega mais informação ao modelo, entretanto traz o problema conhecido como *Maldição da Dimensionalidade* [17].

Uma das maneiras empíricas de identificar um valor ideal para o parâmetro k é observar o decaimentos dos autovalores retornados pela decomposição SVD. Os autovalores representam o peso das respectivas variáveis latentes, logo, como o SVD retorna os autovalores em ordem decrescente, basta observar o ponto de estabilidade do gráfico, ou seja, o ponto onde considerar mais variáveis não traz ganho significativo. Na Figura 2, mostramos o decaimento dos autovalores para o nosso estudo de caso e, como é possível observar, o ponto de estabilidade se dá por volta do 200^{o} autovalor.

F. Testes

Nosso corpus contendo 1000 tweets devidamente rotulados entre as classes *relevante* e *irrelevante* foi dividido em 10 partições distintas, cada uma com 100 tweets. Com a finalidade de termos segurança estatística a respeito da acurácia dos classificadores, em cada uma das possíveis configurações, realizamos a *validação cruzada*.

Em outras palavras, das 10 partições, 9 foram utilizadas para formar o conjunto de treinamento, enquanto que a remanescente constituiu o conjunto de teste. Os tweets do conjunto de treinamento foram representados no espaço vetorial, segundo a devida configuração, e usados para treinar o classificador. Uma vez treinado, este era usado para descobrir as classes dos tweets no conjunto de teste.

Este processo foi repetido 10 vezes, de forma que cada uma das partições serviu como conjunto de teste uma única vez. A acurácia do classificador foi analisada a cada vez que o mesmo foi utilizado no conjunto de teste. Calculamos então a média e o desvio padrão desses resultados, de modo que, nos

gráficos apresentados na seção V, os pontos indicam a média e as barras verticais indicam o desvio padrão da acurácia.

V. Resultados

Na Figura 3, onde comparamos o desempenho do Naive Bayes em relação às possíveis configurações, observa-se nitidamente que o método Booleano favorece o classificador frente aos outros métodos de representação. Além disso, ao considerarmos mais dimensões, percebe-se que o desempenho piora seja qual for o método utilizado.

Já na Figura 4, onde apresentamos os resultados para a Árvore de Decisão, vê-se um comportamento muito instável para todas as configurações. As curvas de desempenho parecem "andar de lado", não sendo possível identificar uma tendência para as mesmas, nem de melhora nem de piora. Na Figura 5, apresentamos o desempenho do classificador de Máxima Entropia, onde, ao contrário do Naive Bayes, percebe-se, com o aumento do numero de dimensões, uma melhora para os métodos TF e TF-IDF e uma piora para o método Booleano.

Quando consideramos o método SVD, vemos na Figura 6 que este favorece a Máxima Entropia, a qual é acompanhada de perto pelo Naive Bayes. Novamente, o desempenho da Árvore de Decisão parece "andar de lado" e, mesmo em seus melhores momentos, este classificador fica muito aquém dos demais.

A própria estrutura do Twitter pode nos ajudar a entender os resultados obtidos. Os tweets são textos com no máximo 140 caracteres, ou seja, o usuário precisa passar sua mensagem em poucas palavras. Além disso, pelo fato de ser popular entre o público jovem, encontramos muitas expressões coloquiais, gírias, *emoticons*, menção a outros usuários, links para páginas Web, etc..

Concluímos então que as opiniões dos usuários são emitidas de maneira compacta e simbólica, muito diferente de uma opinião emitida em um artigo, por exemplo, onde é possível acompanhar o pensamento do autor conforme o mesmo vai desenvolvendo seu argumento. A interpretação dos tweets se assemelha a uma decodificação, i.e., depende fortemente dos elementos presentes no texto.

Desta forma, a presença de uma palavra específica pode fazer total diferença em nossa classificação. Verbos como *adorar*, *gostar* ou *detestar* indicam em si mesmos, independente das demais palavras empregadas, que a probabilidade do autor estar expressando sua opinião é alta. Logo, o método Booleano é o mais adequado para representar os tweets, uma vez que indica claramente a presença ou ausência das palavras mais importantes. Ademais, como é pouco provável que tais palavras se repitam, tendo em vista o tamanho limitado do tweet, buscar uma representação com base na frequência das palavras é uma tentativa ineficaz, segundo o desempenho apresentado por TF e TF-IDF.

A classificação com base no método SVD, por sua vez, requer que as variáveis latentes extraídas sejam representativas o suficiente para indicar a presença ou ausência de uma opinião no texto. Para tal, é necessário um conjunto de treinamento extenso, algo que não conseguimos devido à dificuldade inerente ao processo de rotulagem, i.e., é preciso que alguém leia, interprete e rotule cada tweet. Todavia, vale ressaltar a eficiência do método Booleano frente a este, uma vez que recai

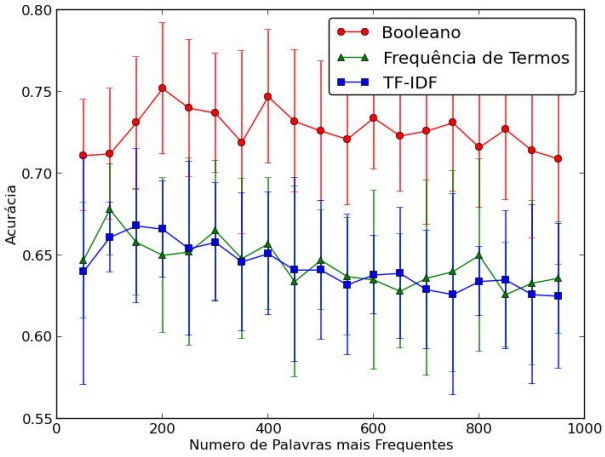

Figura 3. Tweets classificados pelo Naive Bayes

muito mais sobre as palavras escolhidas como dimensões do que sobre o tamanho do conjunto de treinamento.

Por este mesmo motivo, entendemos a diferença de desempenho entre os classificadores. Tanto a Árvore de Decisão quanto a Máxima Entropia dependem muito mais da extensão e da completude do conjunto de treinamento do que o Naive Bayes. O primeiro extrai suas regras de classificação com base nesse conjunto enquanto o segundo extrai suas restrições do mesmo. Se tal conjunto for incompleto ou limitado, isto irá impactar diretamente no desempenho desses classificadores. Ao contrário, o Naive Bayes assume a independência entre as variáveis que representam o texto e seu desempenho está mais atrelado ao comportamento dessas variáveis dentro do conjunto do que na extensão ou completude do mesmo.

Por exemplo, no problema em questão, conforme já mencionado, a presença ou ausência de certas palavras já é suficiente para classificar o tweet como relevante ou irrelevante. Assim, ao utilizarmos o método Booleano, podemos assumir a independência entre as variáveis que representam o texto. Além disso, se o mapeamento entre essas variáveis e as respectivas classes estiver correto no conjunto de treinamento, independente do tamanho do mesmo, o Naive Bayes classificará as instâncias de maneira correta. Vemos, portanto, o motivo deste classificador ter obtido desempenho tão superior aos demais.

VI. Conclusões

Neste trabalho, realizamos um estudo de caso de proporções consideráveis: procuramos filtrar tweets que transmitissem a opinião do usuário a respeito de um candidato à prefeitura da cidade do Rio de Janeiro, durante a campanha eleitoral de 2012.

Atingimos uma acurácia média de até 75% utilizando o método Booleano e o classificador Naives Bayes. Discorremos sobre como este método e este classificador são os mais adequados, dentre os considerados para o trabalho, dada a estrutura do problema. Entretanto, faz-se necessário maiores estudos, com o emprego de classificadores mais robustos, a fim

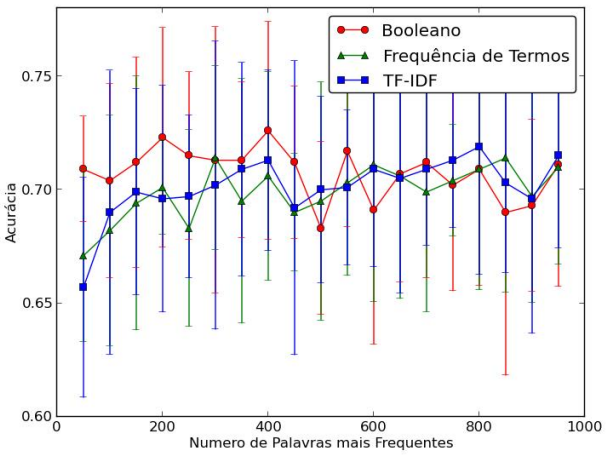

Figura 4. Tweets classificados pela Árvore de Decisão

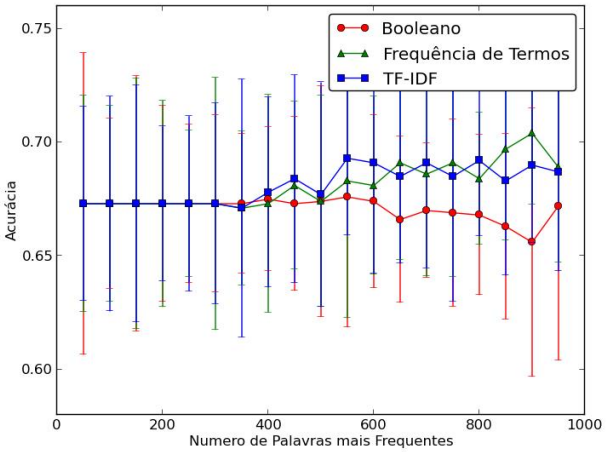

Figura 5. Tweets classificados pela Máxima Entropia

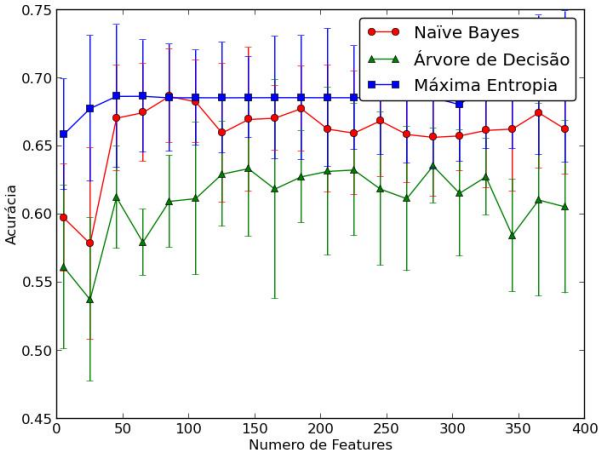

Figura 6. Tweets representados pelo SVD

de alavancar o processo de filtragem de tweets. Vislumbramos que isto será uma forte tendência a medida que o processo eleitoral se torna cada vez mais "informatizado".

Ademais, o tempo necessário para montar um conjunto de treinamento extenso, mesmo com a ajuda de uma aplicação Web, pode levar dias e é altamente suscetível a falhas de rotulagem. Desta forma, desejamos aplicar técnicas de Aprendizado Ativo [18] a fim de tornar a construção de tal conjunto mais eficiente.

REFERÊNCIAS

[1] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in *Proceedings of the Workshop on Languages in Social Media*, ser. LSM '11, 2011, pp. 30–38.

[2] V. N. Khuc, C. Shivade, R. Ramnath, and J. Ramanathan, "Towards building large-scale distributed systems for twitter sentiment analysis," in *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, ser. SAC '12, 2012, pp. 459–464.

[3] F. Bravo-Marquez, M. Mendoza, and B. Poblete, "Combining strengths, emotions and polarities for boosting twitter sentiment analysis," in *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, ser. WISDOM '13, 2013, pp. 1–9.

[4] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010.

[5] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, "A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle," in *Proceedings of the ACL 2012 System Demonstrations*, ser. ACL '12, 2012, pp. 115–120.

[6] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," in *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 2010, pp. 178–185.

[7] A. Jungherr, "Tweets and votes, a special relationship: The 2009 federal election in germany," in *Proceedings of the 2Nd Workshop on Politics, Elections and Data*, ser. PLEAD '13. New York, NY, USA: ACM, 2013, pp. 5–14.

[8] D. Gayo-Avello, "A meta-analysis of state-of-the-art electoral prediction from twitter data," *CoRR*, vol. abs/1206.5851, 2012.

[9] M. W. Berry and M. Browne, *Understanding Search Engines: Mathematical Modeling and Text Retrieval*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1999.

[10] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999.

[11] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2000.

[12] G. Strang, *Introduction to Linear Algebra*. Wellesley, MA: Wellesley-Cambridge Press, 2009.

[13] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000.

[14] M. H. DeGroot, *Probability and Statistics*. Addison-Wesley, 1986.

[15] L. Rokach and O. Maimon, *Data Mining with Decision Trees: Theory and Applications*. River Edge, NJ, USA: World Scientific Publishing Co., Inc., 2008.

[16] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," *Comput. Linguist.*, vol. 22, no. 1, pp. 39–71, Mar. 1996.

[17] M. Köppen, "The curse of dimensionality," in *5th Online World Conference on Soft Computing in Industrial Applications (WSC5)*, 2000, pp. 4–8.

[18] M. B. Pasinato, C. E. Mello, and G. Zimbrão, "Active learning applied to rating elicitation for incentive purposes," in *Advances in Information Retrieval - 37th European Conference on IR Research, ECIR 2015, Vienna, Austria. Proceedings*, 2015, pp. 291–302.