Conectando Opiniões a Opinadores: Um estudo de caso sobre protestos políticos no Brasil*

Ramon Vieira , Alan Neves , Fernando Mourão , Leonardo Rocha

DCOMP/UFSJ - São João del-Rei, MG , Brasil

{ramonv,aneves,fhmourao,lcrocha}@ufsj.edu.br

Abstract. *Sentiment analysis (SA) on Social Media content, as well as the Influential Users Detection (IUD), also called opinion-leaders, provide valuable information for many applications. Despite the intrinsic relation between opinions and opinion-leaders, most of the recent works focus exclusively on one of these two tasks. By empirical assessments on a data sample of tweets about the Brazilian president, this work demonstrates the potential benefits of combining SA Methods with IUD ones. In our analysis, we identified distinct behaviors of opinion propagation and demonstrated that the collective opinion may be accurately estimated by using a few opinion-leaders.*

Resumo. *Análise de Sentimento (AS) sobre conteúdo de Mídias Sociais, bem como a Identificação de Usuários Influentes (IUI), também chamados de opinadores, provêm informações valiosas atualmente. Apesar da intrínseca relação entre opiniões e opinadores, muitos dos trabalhos recentes focam exclusivamente em uma das duas tarefas. Por meio de avaliações empíricas em uma amostra de dados de tweets relacionada à presidente do Brasil, nesse trabalho apresentamos os potenciais benefícios de se combinar métodos de AS com os de IUI. Em nossas análises, identificamos comportamentos distintos de propagação de opiniões e demonstramos que a opinião coletiva pode ser estimada com precisão utilizando tweets relacionados a poucos opinadores.*

1. Introdução

Mídias Sociais vêm se consolidando como um importante ambiente em que pessoas publicam suas opiniões sobre variados assuntos na WEB. Além disso, aplicações de mídias sociais se tornaram decisivas nos processos de tomada de decisão dos usuários. De fato, existe um número crescente de usuários a procura de *reviews* e recomendações sobre produtos e serviços em interações sociais online antes de efetuarem uma escolha [Edelman 2010]. Nesse contexto, compreender e modelar apropriadamente opiniões predominantes de uma dada população (e.g., a opinião coletiva) a respeito de um dado tema (e.g., produto, serviço, etc.), bem como identificar o subconjunto de pessoas formadoras de opinião (opinadores), os quais são capazes de persuadir as outras em relação a um determinado tema, surgiram como problemas relevantes para diversas aplicações, tais como recomendação, publicidade, avaliação de marcas, entre outras.

Apesar dos conceitos de opinião e opinadores apresentarem uma relação intrínseca de difícil dissociação, principalmente quando falamos de dados oriundos de

*Esse trabalho foi parcialmente financiado por CNPq, CAPES, FINEP, Fapemig, e INWEB.

redes sociais, em Ciência da Computação esses dois conceitos vem sendo tratados e abordados distintamente. Grande parte das pesquisas foca em identificar quais são as opiniões preponderantes (i.e. opinião coletiva) ou quais são os principais opinadores responsáveis pela propagação dessas opiniões. Enquanto as opiniões são avaliadas por técnicas de Analise de Sentimento (AS) [Zhao et al. 2012, Rocha et al. 2015], sem necessariamente identificar quem são os responsáveis pelas opiniões analisadas, os opinadores são determinados por técnicas de identificação de usuários influentes (IUI) [Ilyas and Radha 2011, Lee et al. 2010, Silva et al. 2013, Page et al. 1999, Neves et al. 2015], sem que o teor de suas opiniões seja considerado. Uma análise combinada, além de melhorar a identificação adequada tanto das opiniões quanto dos opinadores, pode fornecer maior conhecimento sobre o processo de difusão de informação na Web. Apesar da relevância, encontramos poucos trabalhos na literatura focados em combinar AS e IUI.

Nesse trabalho, apresentamos um estudo preliminar que avalia duas questões principais a respeito da combinação entre AS e IUI. Primeiramente, visamos identificar quais são os principais fatores que podem afetar a eficácia e a relevância da combinação dessas técnicas em cenários reais. Em segundo lugar, pretendemos mostrar alguns potenciais benefícios ao realizar essa combinação. Neste sentido, propomos uma metodologia de quatro passos aplicável a cenários distintos. No primeiro passo, avaliamos as características do domínio que podem afetar AS e IUI, por exemplo, a dinamicidade temporal [Mourão et al. 2008]. O segundo passo refere-se às limitações das técnicas de AS e IUI, por exemplo, sensibilidade quanto a amostra de dados. No terceiro, avaliamos o quão correlacionadas estão AS e IUI. Por fim, propomos duas estratégias diretas para demonstrar o potencial em se combinar AS e IUI. A primeira visa qualificar as opiniões propagadas por cada opinador, enquanto a segunda estratégia pretende estimar a opinião coletiva através de uma análise utilizando apenas alguns opinadores.

Para avaliar a metodologia proposta, conduzimos uma análise empírica em uma amostra de dados do *Twitter*, dada a sua relevância no processo de difusão de informação na Web. Essa amostra corresponde a *tweets* em português relacionados à presidente do Brasil, Dilma Rousseff, postados durante protestos políticos em abril de 2015. Em nossas análises, adotamos as técnicas propostas em [Rocha et al. 2015] e [Neves et al. 2015] para identificar a opinião coletiva e os opinadores, respectivamente. Nossos experimentos indicaram que IUI é mais sensível ao tamanho da amostra do que AS. Além disso, identificamos que técnicas tradicionais de IUI não são capazes de identificar opinadores cujos sentimentos diferem da opinião coletiva. Por outro lado, observamos que a opinião coletiva pode ser estimada com precisão avaliando *posts* de um pequeno número de opinadores. Esses resultados apontam direcionamentos de pesquisa relevantes para a área, evidenciando abordagens novas e promissoras para tratar AS e IUI na Web.

Todas as implementações e execuções de experimentos foram realizadas pelo aluno Ramon Vieira, sob a orientação do professor Leonardo Rocha. A concepção da metodologia bem como as análises de todos os resultados foram feitas em conjunto, aluno e professor, com a colaboração do professor Fernando Mourão. Além disso, esse trabalho contou com o colaboração do aluno Alan Neves, que nos auxiliou na adaptação da técnica de Identificação de Usuários Influentes para o cenário avaliado nesse trabalho.

2. Trabalhos Relacionados

Existem basicamente duas abordagens para técnicas de AS, as supervisionadas e não-supervisionadas. Técnicas supervisionadas normalmente são adaptações de algoritmos de Aprendizado de Máquina tradicionais para aprender as classes relacionadas a sentimento em aplicações de Mídias Sociais. Esses cenários são, em sua maioria, constituídos de textos curtos, assim o desafio é construir modelos apropriados utilizando poucas informações [Brody and Diakopoulos 2011, Zhao et al. 2012, Hu et al. 2013]. Dada a dificuldade em se obter um conjunto de treinamento, métodos não-supervisionados estão assumindo um importante papel na busca de efetivas e eficientes abordagens para AS em conteúdos de mídias sociais. Grande parte desses métodos são baseados em lexicons e possuem dois passos distintos. No primeiro, a consolidação de um lexicon é realizada [Rocha et al. 2015]. Baseado em tal lexicon, o segundo passo foca na identificação do sentimento de cada postagem distinta [O'Connor et al. 2010]. A maioria dos trabalhos existentes focam em análises a nível de documento. O sentimento coletivo é derivado da agregação dos sentimentos de cada documento. Diferentemente, em [Rocha et al. 2015], os autores propõem um novo método que determina o sentimento coletivo diretamente, pela análise de um enorme grafo de termos, construído de acordo com a coocorrência de termos em cada documento.

Técnicas de IUI podem ser divididas em três grupos principais. O primeiro grupo consiste de estratégias que levam em conta a estrutura de redes estabelecidas através do relacionamento entre usuários. Nessa classe temos o trabalho que utiliza o algoritmo de *PageRank*™ para calcular uma pontuação de influência para cada usuário, considerando apenas relacionamentos e a propagação na rede [Page et al. 1999]. Outra estratégia desta classe é o PCC [Ilyas and Radha 2011], o qual usa uma métrica baseada em centralidade para determinar vizinhanças influentes em uma rede. O segundo grupo compreende estratégias que exploram o conteúdo e o fluxo de informação para determinar os opinadores. Neste grupo, estão presentes as estratégias ProfileRank [Silva et al. 2013] e Leitores Efetivos [Lee et al. 2010]. Enquanto o ProfileRank modela a difusão de informação considerando apenas a ordem temporal na qual as mensagens são propagadas em uma rede social, Leitores Efetivos avalia a difusão de informação como um efeito em cascata que tópicos têm sobre usuários. A terceira classe corresponde a estratégias focadas em sumarizações estatísticas de logs de atividade dos usuários. Estas estratégias objetivam determinar uma pontuação de influência para cada usuário de acordo com alguns de seus atributos, tais como número de seguidores e número de *posts* propagados. O principal obstáculo desses trabalhos existentes na literatura é que não há consenso entre eles [Neves et al. 2015].

Não identificamos muitos trabalhos que combinassem AS com IUI. Em [Bigonha et al. 2010] os autores apresentam uma métrica para determinar usuários influentes, baseado em três tipos de informação: a rede de menções, a polaridade de conteúdo dos tweets publicados (os tweets foram classificados individualmente e manualmente), e a qualidade destes tweets. Assim, torna-se possível determinar usuários influentes com viés negativos e positivos. Em [Bae and Lee 2012], é apresentada uma análise relacionando usuários famosos e influentes no Twitter (e.g., Barack Obama e Britney Spears) com suas audiências positivas e negativas, determinando assim seu grau de aprovação. Entretanto, questões tais como se é possível derivar sentimento coletivo pela análise de um pequeno número de opinadores ou quão forte é a correlação entre sentimento coletivo e os opinadores permanecem negligenciada.

3. Metodologia

Nessa seção, apresentamos uma metodologia de quatro passos para quantificar fatores distintos que afetam a análise combinada entre AS e IUI em domínios reais, bem como os potenciais benefícios dessa análise. Basicamente, a metodologia proposta leva em consideração a dinâmica temporal do domínio, a sensibilidade à amostragem dos métodos e a reciprocidade observada entre a opinião coletiva e as opiniões propagadas pelos usuários influentes (opinadores). É importante deixar claro que não estamos assumindo uma avaliação fechada e completa para todas as questões existentes. Nossa proposta é estabelecer direções de pesquisa promissoras para a área.

3.1. Análise de Dinâmica Temporal

A principal característica que afeta a análise combinada entre AS e IUI é a dinâmica temporal inerente ao domínio de análise. A premissa é que opiniões e opinadores temporalmente não alinhados em relação às oscilações observadas não estão diretamente correlacionados. De fato, sempre que a opinião coletiva oscila mais rapidamente do que o subconjunto de opinadores, ou vice-versa, pode não ser possível relacionar ambos. Por exemplo, em cenários nos quais o conjunto de opinadores não está adequadamente consolidado, devido à alta dinamicidade temporal, pode ser desafiador, ou mesmo impraticável, estimar o sentimento propagado por eles. Dessa maneira, esse passo objetiva mensurar o quão dinâmico são as opiniões predominantes e os opinadores em um domínio.

A respeito da opinião coletiva, mensuramos sua dinâmica temporal como segue. Primeiro, derivamos a opinião coletiva O da amostra de dados inteira D, usando um método de AS existente na literatura. Especificamente, adotamos neste trabalho o método SACI [Rocha et al. 2015]. O SACI é relevante para o nosso objetivo uma vez que ele foi originalmente proposto para estimar, de forma eficiente, o sentimento coletivo em amostras de dados, ao invés de agregar o sentimento derivado para cada documento individual. Além disso, os autores demonstraram que o SACI é mais efetivo em estimar a opinião coletiva do que métodos de AS baseados em agregação. O SACI representa O como uma distribuição de probabilidades entre as classes de sentimento positiva, negativa e neutra. Dessa maneira, dividimos D entre unidades temporais de mesmo tamanho (e.g. dias, semanas, meses). Então, estimamos a opinião coletiva O_t usando apenas os *posts* pertencentes a cada unidade temporal distinta t. Finalmente, realizamos uma inspeção visual sobre as distribuições derivadas. Quanto mais dinâmico for um domínio, mais diferentes são as opiniões estimadas em unidades temporais distintas.

Por sua vez, a dinâmica temporal dos opinadores foi medida como segue. Primeiro, identificamos a lista ordenada L de top-k opinadores em D utilizando algum método de IUI. Especificamente, usamos aquele apresentado em [Neves et al. 2015], uma estratégia de *meta-learning* baseada em PCA que combina linearmente informações ortogonais exploradas por distintos métodos de IUI. Denominamos esse método de PCA-IUI. Escolhemos o PCA-IUI uma vez que ele combina estratégias não-consensuais em uma única, capaz de capturar perspectivas distintas. Novamente, consideramos unidades temporais distintas em D e derivamos uma lista distinta L_t de top-k opinadores para cada unidade de tempo t. Em seguida, comparamos cada lista L_t com L usando a correlação generalizada de Kendall's tau

3.2. Análise de Sensibilidade à Amostragem

Além de características do domínio, limitações inerentes aos métodos selecionados para realizar AS coletiva e a IUI podem afetar as nossas análises. Atenção especial deve ser dada para a sensibilidade desses métodos ao tamanho da amostra. É bem conhecido o impacto do tamanho da amostra no processo de aprendizado em áreas distintas [Zhao et al. 2012, Mourão et al. 2008]. A premissa, nesse caso, é que a amostra de dados disponível para análise é suficiente para prover um processo de aprendizagem adequado. Claramente, este requisito depende do método específico usado, bem como da própria tarefa de aprendizado. O desafio, entretanto, é como quantificar tal sensibilidade.

A fim de resolver essa questão, contrastaremos os resultados dos métodos de AS e IUI quando aplicados a amostras aleatórias de tamanhos distintos. Considere novamente a amostra original de dados D. Suponha que comecemos com uma amostra pequena S_i composta por X *posts* distintos aleatoriamente escolhidos de D, onde $X \ll |D|$. Começamos a análise gerando novas amostras S_{i+1} ao adicionar a S_i outros X posts presentes no conjunto $D - S_i$, aleatoriamente escolhidos sem repetição. Em cada amostra S_i aplicamos o SACI e o PCA-IUI contrastando os resultados com aqueles obtidos sob todo conjunto D. Para AS realizamos este contraste ao comparar as distribuições distintas por meio de uma métrica de aproximação de erros, a Distância Euclidiana. Quanto menor a Distância Euclidiana entre duas distribuições, maior a concordância entre elas. Para a IUI, usamos a correlação Kendall's tau, conforme previamente mencionado. Com o objetivo de obtermos robustez estatística, repetimos este processo 10 vezes e consideramos para avaliação os valores de média, bem como o desvio padrão.

3.3. Análise de Reciprocidade

A terceira questão se refere à reciprocidade entre a opinião coletiva de uma população e as opiniões propagadas pelos opinadores. Consideramos que essas duas informações são recíprocas sempre que a opinião coletiva corresponder à opinião predominante entre os opinadores, e vice-versa. Esta é uma informação relevante uma vez que listas top-k de opinadores podem não ser representativas em alguns domínios, emitindo opiniões não absorvidas ou propagadas por toda a população. De forma a realizar essa análise, dividimos a amostra de dados entre três subcoleções disjuntas D_c, de acordo com a classe de sentimento c (i.e., positivo, negativo e neutro) estimado para cada documento. Nesse caso, adaptamos o SACI para realizar análise individual seguindo as sugestões do artigo original. Em seguida, derivamos uma lista L_c de top-k opinadores usando apenas os *posts* e usuários presentes em cada subcoleção D_c. Por fim, comparamos cada uma das listas L_c com os top-k opinadores L derivados de D, usando a correlação Kendall's tau generalizada com penalidade $p = 0$. Quanto maior a correlação entre L_c (relacionada a classe predominante c) e L, mais recíprocas serão a opinião coletiva e a opinião dos opinadores.

3.4. Análise de Ganho de Informação

Por fim, propomos dois experimentos para avaliar os potenciais benefícios de combinar AS com IUI. No primeiro experimento visamos identificar comportamentos distintos dos opinadores em relação aos tipos de opinião comumente propagadas. Nesse sentido, primeiro aplicamos o PCA-IUI na amostra de dados total D, identificando uma lista ordenada L. Em seguida, para cada usuário influente distinto, aplicamos o SACI em seu conjunto de *posts*, derivando uma distribuição de sentimento para esse usuário. Finalmente,

agrupamos esses opinadores de acordo com as suas respectivas distribuições de sentimento. Cada grupo representa um padrão de comportamento distinto. Esse tipo de análise permite qualificar o impacto de opinadores no processo de difusão de informação na WEB.

O segundo experimento objetiva verificar se a opinião de poucos opinadores é suficiente para estimar a opinião coletiva da população inteira. Isso é relevante uma vez que poderia facilitar uma AS mais acurada em muitos domínios em que não é possível se obter uma amostragem representativas da população. Assim, derivamos novamente a lista L de top-k opinadores considerando a amostra de dados completa D. Em seguida, aplicamos o SACI no subconjunto de *posts* publicados somente pelos opinadores identificados, derivando uma distribuição de sentimentos para esse subconjunto. Por fim, calculamos a Distância Euclidiana entre essa distribuição de sentimento e a distribuição de sentimento derivada para D. Repetimos esse processo considerando valores distintos de k.

4. Estudo de caso

Essa seção visa validar a metodologia proposta e apontar evidências da relevância deste estudo em domínios reais. Nesse sentido, aplicamos a metodologia em uma amostra de dados coletada do *Twitter*. Iniciamos a discussão descrevendo a amostra de dados coletada bem como as configurações do ambiente de execução. Em seguida, apresentamos os resultados relacionados a cada passo da metodologia, juntamente com as principais conclusões e implicações para a área.

4.1. Coleção de Dados e Configurações do Ambiente

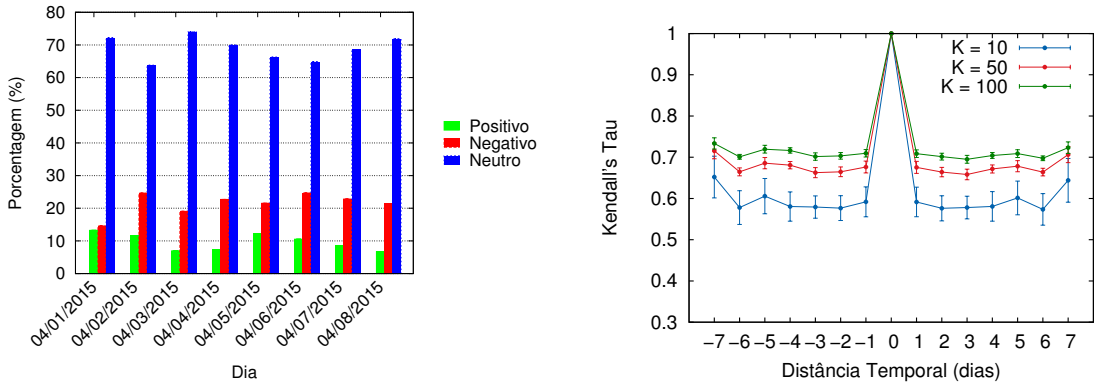
Utilizamos uma amostra de dados coletada do *Twitter*, que corresponde a *tweets* relacionado aos protestos políticos em 2015 em relação a presidente do Brasil, Dilma Rousseff. A coleta foi feita por meio da *API de Streaming* do Twitter utilizando as palavras-chave 'Dilma' e 'Dilma Rousseff'. Removemos de cada tweet a pontuação, caracteres especiais, padrões repetitivos, bem como convertemos todas as letras para minúsculo. Além disso, removemos palavras com menos de três letras ou mais de 12 letras, URLs e menções a usuários do Twitter. Esses dados foram coletados entre **01-04-2015 e 08-04-2015** e contém **225,885 tweets**. Utilizando a API REST coletamos também informações sobre os **61,294 usuários** que publicaram os tweets coletados, tais como seguidores e amigos.

Conforme mencionado, em nossos experimentos adotamos para AS o algoritmo SACI [Rocha et al. 2015]. SACI tem quatro parâmetros relacionados à construção do *lexicon*, *suporte mínimo, confiança, número de sementes e distância máxima de propagação*, que foram fixados como 3, 0.8, 100 e 3, respectivamente. No passo de análise, o SACI tem o parâmetro *raio máximo de transformação*, definido como 4. Todos esses valores correspondem às configurações com os melhores resultados reportados pelos autores. No que se refere a IUI, adotamos a PCA-IUI [Neves et al. 2015], que corresponde a uma estratégia de *meta-learning* baseada em PCA que combina sete estratégias: PCC (número máximo de autovetores=100), Pagerank (erro máximo entre iterações consecutivas=0.0000001 e p=0.85), Profilerank (erro máximo entre iterações consecutivas=0.0000001), Closeness, Betweenness, Número de Retweets e Número de Seguidores.

4.2. Análise de Dinâmica Temporal

A fim de se avaliar a dinâmica temporal em nossa amostra de dados, a dividimos em dias, definindo oito subcoleções disjuntas. Começando as nossas análises pela Análise

de Sentimento, a figura 1 (a) retrata as distribuições de sentimento derivadas pelo SACI ao longo dos dias. Observamos distribuições bastante similares ao longo dos dias, com a classe neutra apresentando cerca de 65%-70% de porcentagem de ocorrência, seguido das classes negativa e positiva, respectivamente. Uma exceção é observada em 1^o de abril, quando a porcentagem de *posts* negativos foi quase a mesma da porcentagem de positivos. A explicação para esse comportamento é que se trata do Dia da Mentira, resultando em muitos *posts* irônicos. Além disso, ao aplicarmos o SACI na amostra de dados completa, descobrimos que a distribuição de sentimento é de 63.70% de *posts* neutros, 23.91% de negativos e 12.39% de positivos. Esse resultado mostra que o sentimento coletivo estimado para qualquer dia é similar ao sentimento de todo o período de tempo avaliado, o que retrata o domínio avaliado como estável em relação à opinião coletiva.



(a) Opinião Coletiva Estimada ao longo do Tempo. (b) Opinadores Identificados ao longo do Tempo.

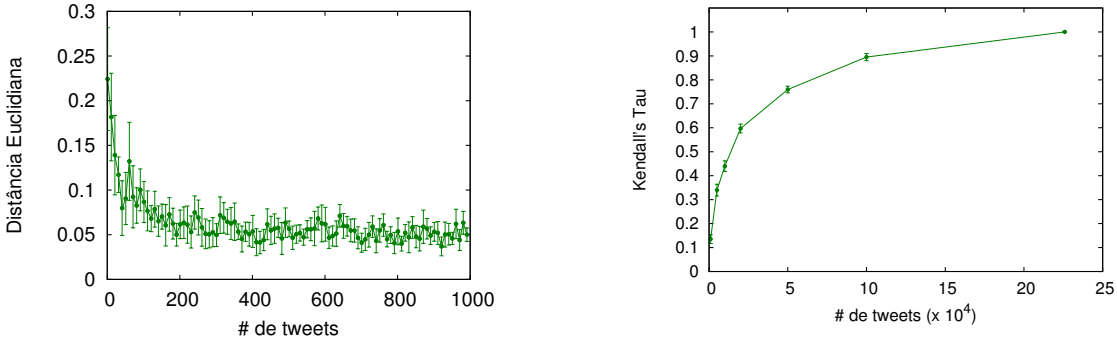
Figura 1. Dinâmica Temporal da amostra de dados coletadas a partir de Twitter.

A figura 1 (b) apresenta a dinâmica temporal dos opinadores. Neste experimento, avaliamos a similaridade dos top-k opinadores ao longo do tempo considerando valores distintos de k (e.g., $k = 10$, $k = 50$ e $k = 100$). Também plotamos os valores de Kendall's tau entre cada par de dias t_i e t_j usando a distância temporal entre eles (i.e., a distância é igual a $t_j - t_i$). Cada ponto neste gráfico representa, portanto, o valor de correlação média para todos os pares de dias cuja distância temporal é a mesma. Primeiramente, observamos que listas top-k pertencentes a dias distintos apresentam valores de correlação acima de 60% e quase inalterados conforme a distância temporal aumenta. Isto demonstra que este domínio também é estável em relação ao conjunto de opinadores ao longo do tempo, apresentando poucas alterações nas listas. Também observamos que quanto maior o tamanho das listas, maior é a correlação média das listas top-k distintas. Conforme esperado, pequenas alterações em listas pequenas de opinadores produzem grandes impactos nas correlações inter-listas ao longo do tempo, tal como observado para $k = 10$.

4.3. Análise de Sensibilidade à Amostragem

Em relação a sensibilidade dos métodos quanto à amostragem, as figuras 2 (a) e (b) apresentam os resultados para as tarefas de AS e IUI, respectivamente. Novamente, os valores plotados correspondem à média de 10 execuções. Observamos que, na amostra de dados avaliada, AS foi menos sensível ao tamanho da amostra do que a IUI. De fato, foi possível aproximar a opinião coletiva usando apenas 200 *tweets* (distância euclidiana menor que 0,05). Isso significa que usando apenas 0,08% da coleção foi possível aproximar a distribuição de sentimento alcançada usando a coleção inteira, superestimando ou subestimando a classe majoritária por 5% apenas. Por outro lado,

foi necessário considerar grandes amostras para identificar opinadores. Por exemplo, atingimos valores de correlação maiores do que 0,7 usando pelo menos 50.000 *tweets* (22% da amostra de dados). Esses resultados apontam que amostras pequenas de dados podem não ser suficientes para identificar efetivamente opinadores em alguns domínios.



(a) Opinião Coletiva Estimada (b) Opinadores Identificados.

Figura 2. Sensibilidade a Amostragem do SACI (a) e PCA-IUI (b).

4.4. Análise de Reciprocidade

Avaliamos a reciprocidade entre a opinião coletiva e a opinião predominante dos opinadores variando o valor de k nas listas top-k de 5 até 50, tal como mostrado na figura 3. Nela podemos ver que a maioria dos top-k opinadores da amostra de dados total também são os opinadores que propagam a opinião coletiva predominante (a classe neutra, no caso), exibindo correlações inter-listas de cerca de 85% desconsiderando o valor de k. Por outro lado, opinadores que propaguem sentimento positivo em nossas amostras possuem o menor grau de correlação com os top-k opinadores na coleção, evidenciando uma baixa interseção dessas duas listas. Assim, a opinião predominante propagada pelos top-k opinadores em nossos dados é igual a opinião coletiva predominante.

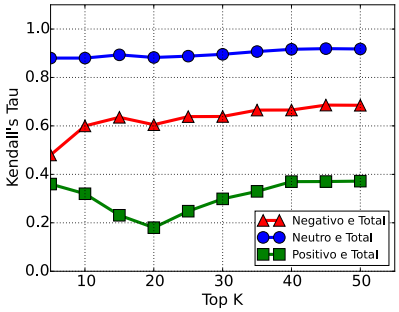
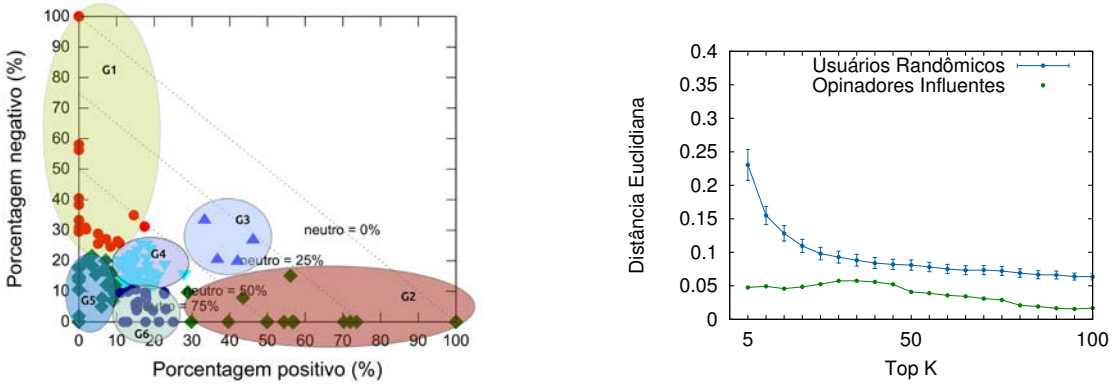


Figura 3. Análise de Reciprocidade entre SACI e PCA-IUI.

Esses resultados nos mostram que métodos de IUI estado-da-arte podem não identificar corretamente usuários individuais que agem como opinadores. Como a identificação de opinadores é enviesada para a opinião predominante e tal opinião é a classe neutra em vários domínios, os opinadores identificados são basicamente jornais ou outras mídias de comunicação. Além disso, é quase impossível identificar opinadores que propaguem opiniões que divirjam da predominante. Por exemplo, apesar de uma companhia estar interessada em identificar os opinadores que não gostaram de um determinado produto, os métodos de IUI atuais podem não ser capaz de encontrá-los. Filtrar *posts* negativos e aplicar os métodos de IUI nos dados resultantes pode não funcionar devido à sensibilidade à amostragem.

4.5. Análises de Ganhos da Combinação de AS e IUI

Por fim, avaliamos os potenciais benefícios de se combinar AS e IUI em nossa amostra de dados. Primeiramente, pretendemos identificar comportamentos distintos de opinadores em relação aos tipos de opinião geralmente propagadas por eles. A figura 4 (a) apresenta os resultados relacionados a este experimento. Para esse gráfico, usamos os top-200 opinadores identificados pelo PCA-IUI. Identificamos seis grupos com comportamentos distintos. O grupo 1 (G1) corresponde a usuários com uma posição negativa bem definida em relação à presidente do Brasil, enquanto o grupo 2 (G2) compreende usuários com uma posição positiva. O grupo 3 (G3) é composto por usuários que apoiam a presidente em alguns aspectos mas não em outros. Por sua vez, os grupos 4, 5 e 6 (G4, G5 e G6) representam usuários neutros, que publicam fatos em sua maioria, ora com uma tendência negativa (G5) e ora positiva (G6). Além disso, esses resultados mostram que AS pode melhorar a IUI, ao extrair informações úteis para o entendimento de como as opiniões se propagam.



(a) Comportamento de Opinadores. (b) Opinião Coletiva Estimada.

Figura 4. Análises dos benefícios de combinar AS e IUI.

No segundo experimento, verificamos se a opinião de poucos opinadores é suficiente para estimar a opinião coletiva de uma população inteira. A figura 4 (b) apresenta os resultados quando variamos os top-k opinadores usados no experimento entre $k = 5$ e $k = 100$. O gráfico também mostra os resultados do mesmo experimento quando usamos k usuários aleatoriamente escolhidos da amostra de dados. Cada ponto i da curva representa a distância euclidiana média da posição i em 100 execuções, bem como o desvio padrão. Observamos que a opinião definida pelos top-k opinadores está mais próxima da opinião coletiva do que a opinião relacionada a usuários aleatórios. Observe que usar poucos opinadores (e.g., os top-5) é suficiente para produzir uma aproximação melhor do que 100 usuários aleatórios. Além disso, o erro de aproximação (i.e., a distância euclidiana) convergiu rapidamente para ambos conjuntos de usuários. Esses resultados demonstram como técnicas de AS podem se beneficiar dos resultados da IUI.

5. Conclusões e trabalhos futuros

Nesse trabalho apresentamos uma metodologia para quantificar fatores que possam afetar avaliações combinadas de Análise de Sentimento (AS) coletiva e Identificação de Usuários Influentes (IUI) em domínios reais, bem como seus potenciais benefícios. Avaliamos nossa metodologia utilizando postagens relacionadas aos protestos políticos de abril de 2015. Em nossas avaliações identificamos um cenário de estabilidade temporal tanto para as opiniões coletivas quanto para opinadores e que o método de

IUI selecionado apresenta uma maior sensibilidade à amostra do que o método de AS utilizado. Observamos também que a opinião coletiva é recíproca à opinião predominante definida pelos opinadores mais influentes em nossa coleção. Por fim, mostramos que é possível qualificar opinadores identificando opiniões distintas sendo propagadas e que a opinião coletiva pode ser estimada com precisão através de poucos top-k opinadores. Como trabalhos futuros, nossa meta é propor estratégias mais robustas para combinar IUI e AS, bem como aplicar a metodologia proposta em outros domínios.

Referências

Bae, Y. and Lee, H. (2012). Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers. *Journal of the American Society for Information Science and Technology*, 63(12):2521–2535.

Bigonha, C., Cardoso, T. N., Moro, M. M., Almeida, V. A., and Gonçalves, M. A. (2010). Detecting evangelists and detractors on twitter. In *18th Brazilian Symposium on Multimedia and the Web*, pages 107–114.

Brody, S. and Diakopoulos, N. (2011). Cooooooooooooollllllllllll!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In *Proc. of EMNLP 2011*, UK.

Edelman, D. (2010). Branding in the digital age. *Harvard Business Review*, 88:14–18.

Hu, X., Tang, J., Gao, H., and Liu, H. (2013). Unsupervised sentiment analysis with emotional signals. In *Proc. of WWW '13*, pages 607–618, Geneva, Switzerland.

Ilyas, M. U. and Radha, H. (2011). Identifying influential nodes in online social networks using principal component centrality. In *IEEE Int. Communications Conference*.

Lee, C., Kwak, H., Park, H., and Moon, S. (2010). Finding influentials based on the temporal order of information adoption in twitter. In *Proc. of WWW*. ACM.

Mourão, F., Rocha, L., Araújo, R., Couto, T., Gonçalves, M., and Meira Jr., W. (2008). Understanding temporal aspects in document classification. In *WSDM 2008*, USA.

Neves, A., Vieira, R., Mourão, F., and Rocha, L. (2015). Quantifying complementarity among strategies for influeners´ detection on twitter. ICCS.

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

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical report.

Rocha, L., Agostinho, R., Sá, G., Silveira, T., Teixeira, F., Gomes, R., Ferreira, R., and Mourão, F. (2015). Saci: Sentiment analysis by collective inspection on social media content. *Journal of Web Semantics*.

Silva, A., Guimarães, S., Meira, Jr., W., and Zaki, M. (2013). Profilerank: Finding relevant content and influential users based on information diffusion. In *Proc. of the ACM Workshop on Social Network Mining and Analysis*.

Zhao, J., Dong, L., Wu, J., and Xu, K. (2012). Moodlens: an emoticon-based sentiment analysis system for chinese tweets. In *Proc. of KDD 2012*, pages 1528–1531, China.