SACI: Sentiment Analysis by Collective Inspection on Social Media Content*

Rodrigo Chaves[1], Giovanni Sá[1], Ramon Vieira[1]
Fernando Mourão[2], Leonardo Rocha[1]

[1] DCOMP/UFSJ - São João del-Rei, MG , Brasil

[2]DCC/UFMG - Belo Horizonte, MG , Brasil

{rachaves,giovannisa,ramonv,lcrocha}@ufsj.edu.br,{fhmourao}@dcc.ufmg.br

Abstract. *Sentiment analysis on Social Media content represent a valuable information nowadays. Current studies assume that to determine a collective opinion requires the prior knowledge of each individual one. In this paper we present SACI (Sentiment Analysis by Collective Inspection), a lexicon-based unsupervised method that extracts collective sentiments without concerning about individual classifications. We demonstrate that SACI fulfills simultaneously requirements of efficacy and efficiency. Its collective analysis outperforms individual one w.r.t. approximating the collection sentiment. Further, a* Web *tool for real-time sentiment analysis built on the top of SACI demonstrates its usefulness.*

Resumo. *Análise de Sentimento sobre conteúdo de Mídias Sociais representa uma informação valiosa atualmente. Estudos recentes assumem que para determinar uma opinião coletiva necessita-se conhecer as opiniões individuais. Nesse trabalho apresentamos o SACI (Sentiment Analysis by Collective Inspection), um método não-supervisionado baseado em lexicon que extrai sentimento coletivo sem considerar classificações individuais. Demonstramos que o SACI atende tanto a requisitos de eficácia quanto de eficiência, visto que sua análise coletiva permite uma melhor aproximação da opinião coletiva de um conjunto de documentos. Consolidamos esse trabalho em uma ferramenta* Web *de análises de sentimento em tempo real.*

1. Introdução

As opiniões postadas em Mídias Sociais vem se consolidando como uma importante fonte de informação para modelagem e entendimento do comportamento de usuários. Diversas técnicas de análise computacional vem sendo adaptadas para esse novo cenário, com ênfase particular à **Análise de Sentimento** (AS), que consiste em extrair automaticamente conteúdo subjetivo expresso nos dados textuais [Pak and Paroubek 2010, Wilson et al. 2005, Dave et al. 2003, Pang et al. 2012]. Entretanto, AS no contexto de Mídias Sociais é mais desafiador que em outros cenários, como as tradicionais análises de produtos, uma vez que seu conteúdo é composto por um fluxo contínuo de textos curtos, desestruturados e de domínio específico [Dave et al. 2003].

Nesse trabalho focamos no desafio de identificar o sentimento coletivo sobre um determinado alvo (ex: produto, pessoa ou serviço) mencionado em um *stream*

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

de textos, considerando como sentimento coletivo aquilo que é expressado em todo um conjunto de documentos. Estudos recentes assumem que a identificação de opinião coletiva requer um conhecimento a priori de cada opinião individual. Entretanto, classificar textos pequenos de forma individual dificulta a assimilação de informações relativas ao contexto. Sendo assim, propomos o *SACI* (*Sentiment Analysis by Collective Inspection*), um novo método não-supervisionado baseado em *lexicon* que adapta-se de forma eficiente a um grande volume de documentos extraindo o sentimento coletivo sem realizar classificações individuais.

O *SACI* é baseado em um grafo probabilístico entre os termos de um conjunto de documentos e em uma classificação contextualizada a priori desses termos com relação a sua função semântica na consolidação de opiniões (*lexicon*). Caminhos representam subconjuntos de sentenças no grafo e a opinião coletiva é definida percorrendo esses caminhos e aplicando sucessivas transformações semânticas, de acordo com a classe de cada termo alcançado. Como cada caminho tem uma probabilidade de ocorrência, o sentimento coletivo é dado pela soma das probabilidades de ocorrência associadas aos caminhos positivos, negativos e neutros. Nossa hipótese é que a sobreposição por meio de distintos documentos pode enfatizar opiniões consensuais, onde pontos de vista individuais e pouco frequentes se tornam menos relevante para o sentimento coletivo.

Para avaliar o *SACI*, comparamos sua análise coletiva com estratégias agregadas derivadas de dois métodos não-supervisionadas [Pak and Paroubek 2010, Wilson et al. 2005] e dois supervisionados [Dave et al. 2003, Pang et al. 2012] bem estabelecidos na literatura. Além de demonstrar que a análise coletiva supera as individuais em relação a aproximação com as distribuições de opiniões (ganho de até 86%), nossas avaliações comprovam que boas classificações individuais não garantem uma boa análise coletiva e vice-versa. Além disso, demonstramos que o *SACI* satisfaz simultaneamente requisitos de eficácia e eficiência, sendo capaz de lidar com a alta dinamicidade encontrada em cenários de grande demanda, como as Mídias Sociais. De fato, a consolidação de uma ferramenta Web para análise de sentimento de *tweets* (textos publicados no microblog *twitter.com*) baseada no *SACI* enfatiza a utilidade desse trabalho.

A implementação do SACI e as execuções dos experimentos foram realizadas pelo aluno Rodrigo Chaves, sob a orientação do professor Leonardo Rocha. Toda a concepção do algoritmo bem como as análises de resultados foram feitas em conjunto, aluno e professor, com a colaboração do aluno de doutorado do Programa de Pós-Graduação do DCC/UFMG Fernando Mourão. Além disso, esse trabalho contou com o auxílio do aluno Giovanni de Sá, responsável pelo concepção do lexicon utilizado nesse trabalho, e do aluno Ramon Vieira, que colaborou na construção da ferramenta Web *de análise de sentimento em tempo real.*

2. Trabalhos Relacionados

Observamos nos últimos anos um grande interesse na AS de informações publicadas em Mídias Sociais por se tratar de uma rica e enorme fonte de informação subjetiva sobre os usuários. Trata-se de um cenário desafiador, caracterizado por textos curtos, com poucas informações (afetando a eficácia) e grandes volumes de dados (afetando a eficiência). Nessa seção apresentamos uma breve revisão da literatura de

trabalhos que apresentam métodos de AS. Focamos em duas abordagens principais: as estratégias supervisionadas e as não-supervisionadas.

Estratégias supervisionadas focam na adaptação de algoritmos de aprendizado de máquina para que os mesmos considerem informações adicionais, extraídas por métodos de processamento de linguagem natural (ex: classes sintáticas de termos), para tentar suprir as poucas informações contidas nos textos. Pang, Lee e Vaithyanathan [Pang et al. 2012] apresentam uma proposta na qual sequências de termos que ocorrem nos documentos (n-gramas) são utilizadas por um classificador SVM (***S**upport **V**ector **M**achine*) para classificar sentimento em textos. Dave, Lawrence e Pennock [Dave et al. 2003], também apresentam um modelo de classificação supervisionada que atribui valores de -1 a 1 a cada n-grama observado no conjunto de treino, de acordo com sua frequência nos documentos previamente classificados como positivo ou negativo. O sentimento de cada documento é definido como o sentimento da classe com maior soma dos valores dos n-gramas. Essas técnicas, apesar de eficazes, apresentam um custo computacional elevado em função da extração das novas informações e da própria geração dos modelos, tornando-as pouco atrativas para cenários de Mídias Sociais. Assim, os métodos não-supervisionados estão assumindo um importante papel na busca de abordagens eficientes para AS nesses cenários.

A maioria das estratégias não-supervisionadas são baseadas em *lexicons*, estruturas que guardam a função semântica de termos no processo de consolidação de opiniões, e possuem geralmente duas etapas distintas. A primeira é a construção do *lexicon* e a segunda etapa consiste em identificar o sentimento dos documentos baseados nesse *lexicon*. Com relação a construção dos *lexicons*, a maior parte dos esforços podem ser divididos em três categorias principais [Hu et al. 2013]: (1) baseados em anotação humana, em que os termos são classificados manualmente [Wilson et al. 2005] (bons *lexicons* mas de construção trabalhosa); (2) baseados em dicionário que identificam o sentimento de um termo por meio de termos semanticamente relacionados em um dicionário específico (por exemplo, Wordnet) [Baccianella et al. 2010] (computacionalmente eficientes mas não contextualizados); e (3) conhecidos como *corpus-based* nos quais o sentimento dos termos é definido pelo seu contexto, considerando relações entre termos [Pak and Paroubek 2010] (*lexicons* tão bons quanto os construídos manualmente e reduzido custo computacional). O *SACI* é um método não-supervisionado que utiliza um *lexicon* relacionado a essa terceira categoria, o LEGi [Sá et al. 2014].

3. SACI

Conforme mencionamos anteriormente, o *SACI* é um método de análise de sentimento não-supervisionado baseado em *lexicon*, uma estrutura que define a função semântica dos termos na consolidação das opiniões. Dessa forma, a primeira tarefa é definir o *lexicon* que será utilizado. A partir desse *lexicon*, o *SACI* executa a análise sentimento por meio de duas etapas: (1) construção do grafo de transição entre termos; e (2) percurso no grafo inferindo o sentimento coletivo.

3.1. Definição do Lexicon

A escolha adequada do *lexicon* a ser utilizada é crucial para métodos de AS baseado em *lexicons*, como no caso do *SACI*. Conforme mencionado na seção 2, existem

várias estratégias na literatura para a consolidação de *lexicons* e estratégias que definem o sentimento dos termos de acordo com o contexto têm se mostrado mais eficazes. Recentemente, publicamos uma nova proposta de consolidação de *lexicon*, o LEGI [Sá et al. 2014], a qual se mostrou bastante promissora superando significativamente diversas outras propostas da literatura, sendo portanto utilizada no *SACI*. O LEGI assume seis classes semânticas distintas, conforme apresentado na tabela 1, e tais classes são utilizadas pelo *SACI*, conforme descrito a seguir.

Classe	Transformação	Exemplos
Positivo	Torna positiva a polaridade de uma sentença.	bom, legal, fantástico
Negativo	Torna negativa a polaridade de uma sentença.	terrível, feio, ruim
N**E**utro	Não altera a polaridade de uma sentença.	a maioria dos substantivos
Inversor	Inverte a polaridade de uma sentença.	não, nunca
Redutor	Atenua a polaridade de uma sentença.	pouco, pequeno, levemente
Amplificador	Intensifica a polaridade de uma sentença.	muito, grande, extremamente

Tabela 1. Classes polares do vocabulário Português.

3.2. Construção do Grafo

O *SACI* utiliza um grafo para modelar relacionamentos entre pares de termos, o qual é construído baseando-se em caminhos (i.e., sequências de termos) extraídos dos documentos analisados. Define-se o conjunto de entidades avaliadas como nós centrais e partindo-se de cada nó central construímos um caminho extraíndo alguns termos predecessores e sucessores, pertencentes a uma mesma sentença, preservando-se a ordem de ocorrência dos termos. A medida que esses caminhos são extraídos, o grafo de termos é construído como um grafo probabilístico direcionado, modelado como caminhos markovianos tal como em um *Customer Behavior Model Graph* (CBMG)[Mark and Csaba 2007].

Mais formalmente, considere D o conjunto de documentos a serem analisados, M o conjunto de nós centrais relevantes e um parâmetro l (raio máximo) que representa o raio de aplicabilidade de cada transformação de sentimento dentro de cada sentença. Inicialmente, extraímos de cada documento $d_{i'} \in D$ um conjunto $F_{i'}$ de sentenças presentes em $d_{i'}$. Para cada ocorrência de um nó central $m_{j'} \in M$ em uma sentença $f_{k'} \in F_{i'}$, extrai-se um caminho $c_{k'}$ composto pelos x termos predecessores de $m_{j'}$, o próprio nó central $m_{j'}$, e os y termos sucessores de $m_{j'}$, para $x, y <= l$, preservando a ordem de ocorrência dos termos em $f_{k'}$. Termos com menos de três caracteres são ignorados neste passo, uma vez que eles usualmente compreendem artigos e preposições, que não ajudam a inferir semântica. Em seguida, definimos um conjunto C de todos os caminhos extraídos de todas as sentenças em todo o conjunto de documentos. Considerando T como o conjunto de termos distintos observados em D, definimos uma rede G na qual cada termo $t_i \in T$ corresponde a um vértice, e para cada par de termos $t_i, t_j \in T$, existe uma aresta direcionada e ponderada de t_i para t_j se t_i precede t_j em pelo menos um caminho $c_{k'} \in C$. O peso da aresta é definido como a probabilidade de t_i preceder t_j no conjunto C inteiro. Definimos ainda os estados *Entry* e *Exit*, equivalentes aos vértices de entrada e saída do nosso grafo. Se um termo t_i inicia um caminho, irá existir uma aresta do vértice *Entry* para t_i. Da mesma forma, sempre que um termo t_i encerra um caminho, ele será conectado ao vértice *Exit*. Assim, cada caminho pode representar mais de uma sentença de $F_{i'}$. Cada nó t_i de G possui também um atributo que determina classe de sentimento (definido pelo *lexicon*).

É importante notar que o tamanho do grafo é linear quanto ao número de termos distintos, mantendo-se pequeno. Além disso, o mesmo pode ser atualizado para refletir as opiniões mais recentes ou emitidas em um período de tempo específico.

3.3. Análise de Sentimento Coletiva

A análise de sentimento coletiva realizada pelo *SACI* consiste em navegar pelo grafo construído no passo anterior, percorrendo-se todos os caminhos que pertençam ao conjunto C, definindo o sentimento de cada caminho individualmente. Nossa premissa é que a sobreposição entre documentos distintos possa enfatizar opiniões consensuais ou mais frequentes, enquanto pontos de vista individuais são suavizados.

A definição do sentimento de cada caminho no *SACI* é feita por meio de um modelo de transição de estados definido por uma Máquina de *Mealy*, na qual o próximo estado depende do estado atual e do dado de entrada. Uma Máquina de *Mealy* é uma 6-tupla $(F, F_0, \Sigma, \Lambda, T_S, T_O)$, onde F é um conjunto finito de estados, F_0 é o estado inicial, Σ é o alfabeto de entrada, Λ é o alfabeto de saída, T_S é a função de transição que mapeia cada par de estados e um símbolo de entrada para um estado de saída, e T_O é a função de saída que mapeia cada par de estados e um símbolo de entrada para um símbolo de saída. Em nosso caso, definimos $F = \{P, N, E, I, R, A\}$, que corresponde as classes semânticas, $F_0 = \{E\}$ uma vez que o sentimento inicial de qualquer caminho é definida como nula, $\Sigma = \{p, n, e, i, r, a\}$ é a classe do próximo termo do caminho no grafo de transição, $\Lambda = \{p, n, e, \varepsilon\}$ é o sentimento do caminho, T_S e T_O são representados na figura 1. Cada combinação do estado atual e dado de entrada resulta em um símbolo de saída, que representa o novo sentimento do caminho, bem como em um estado de saída. O caminho é sequencialmente percorrido e a cada termo alcançado, o sentimento do caminho é alterado de acordo com regras de transformação definidas pela Máquina de *Mealy*. Este processo continua até que o último termo do caminho seja alcançado. O último símbolo de saída diferente de ε corresponde ao sentimento associado a cada caminho $c_{k'}$ ($sentiment[c_{k'}]$).

Este processo de atualização do sentimento por meio de regras de transformação é sempre feito pela modificação do sentimento do caminho atual, de acordo com a classe semântica do próximo termo t_i. Uma exceção ocorre quando a classe semântica de t_i é *Inversora*. Na comunicação humana, um termo inversor, como "não", em geral, inverte o sentimento de termos sucessores. Assim, deve-se inverter o sentimento resultante dado pelo primeiro termo que sucede t_i no caminho. Outro ponto a ser enfatizado corresponde as classes semânticas *Amplificadora* e *Redutora*. Como podemos ver na figura 1, elas não são usadas no processo de atualização do sentimento dos caminho. Atualmente, o *SACI* usa estas classes para evitar que alguns termos sejam erroneamente classificados como positivos ou negativos.

Além de seu sentimento, cada caminho possui uma probabilidade de ocorrência, tal como em um CBMG. Mais formalmente, cada caminho $c_{k'} = \{Entry, t_1, t_2, \ldots, t_k, Exit\}$, do conjunto de caminhos C, define uma probabilidade $prob[c_{k'}]$, como mostrado pela equação 1, onde $prob[t_j \mid t_i]$ corresponde a probabilidade de se alcançar t_j partindo de t_i e $prob[C]$ denota a soma das probabilidades de todos os caminhos de C. A soma de logaritmos é utilizada para evitar que as probabilidades de caminhos longos tornem-se rapidamente próximas de zero. A figura 2 (a) ilustra o processo de construção da rede para um pequeno conjunto de frases de uma base de críticas de filmes e a figura (b) mostra o cálculo da probabilidade para um caminho.

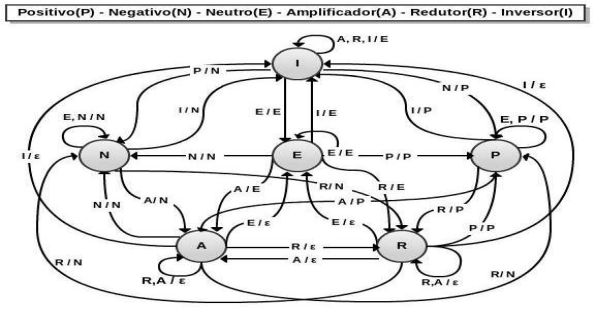
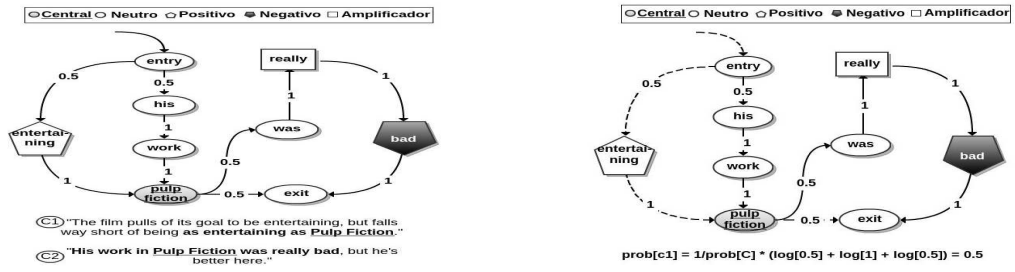


Figura 1. Regras de Transformação (T_S e T_O) definidas pela Máquina de *Mealy*.

$$
\begin{aligned}
prob[c_{k'}] &= \frac{1}{prob[C]} * [\log(prob[t_1|Entry] * prob[t_2|t_1] * \ldots * prob[t_k|t_{k-1}] * prob[Exit|t_k])] \\
&= \frac{1}{prob[C]} * [\log(prob[t_1|Entry]) + \log(prob[t_2|t_1]) + \ldots \\
&\quad \ldots + \log(prob[t_k|t_{k-1}] + \log(prob[Exit|t_k])] \qquad (1)
\end{aligned}
$$

$$
prob[class] = \sum_{\{c_i \in C | sentiment[c_i] = class\}} prob[c_i] \qquad (2)
$$

Utilizando o sentimento relacionado a cada caminho distinto $c_{k'} \in C$ ($sentiment[c_{k'}]$) e sua probabilidade ($prob[c_{k'}]$), definimos a probabilidade de ocorrência relacionada a cada sentimento coletivo $class \in \{neutro, positivo, negativo\}$ no conjunto D conforme mostrado na equação 2. Assim, caminhos incomuns no grafo são menos relevantes para o processo de análise, enquanto caminhos com maior probabilidade são priorizados, suavizando erros causados por caminhos individuais pouco usuais. É importante notar que o *SACI* não requer uma etapa de treinamento, reduzindo significativamente seu custo computacional.



(a) Construção de uma rede de termos com $l = 2$

(b) Probabilidade do caminho C1 (linhas pontilhadas)

Figura 2. Exemplo de Utilização da Rede

4. Avaliação Experimental

Nessa seção apresentamos os resultados relacionados a avaliação do *SACI*. Primeiramente, descrevemos as bases de dados utilizadas, bem como a configuração dos experimentos e parâmetros de execução. Por fim, discutimos os resultados alcançados.

4.1. Bases de Dados

A análise de sentimento coletivo realizada pelo *SACI* depende essencialmente da existência de sobreposição entre conteúdos publicados em mídias sociais. Nossa hipótese é que tal sobreposição é afetada pela localidade temporal destes conteúdos, uma vez

que conteúdos publicados com localidade temporal próxima possuem uma quantidade maior de assuntos correlacionados, gerando um aumento na sobreposição de assuntos, levando a uma análise mais precisa. Visando verificar tal hipótese, optamos por não utilizar bases disponíveis na literatura, uma vez que tais bases não proveem informações sobre essa localidade temporal. Assim, consolidamos dois conjuntos reais de documentos, coletados a partir do *Twitter*, no qual variamos essa característica. O primeiro conjunto é composto de *tweets* relacionados a 20 séries de TV americana de audiência global, como *The Big Bang Theory* e *Two and a Half Men*, entre outros. Nesse conjunto, a coleta foi realizada entre 05/03/2012 e 17/03/2012, mantendo a localidade temporal dos assuntos discutidos. O segundo conjunto é composto por *tweets* relacionados à eleição presidencial americana de 2012 coletados em intervalos distintos durante o ano de 2012. Cada *tweet* de ambas bases foram manualmente classificados em uma das três classes polares (Positivo, Negativo e Neutro) por cinco pessoas diferentes, e cada documento foi assinalado à classe com maior quantidade de votos. Nessa classificação, os avaliadores foram instruídos a identificar se o texto tinha intenção de exaltar, criticar ou simplesmente relatar um fato sobre o alvo analisado. A tabela 2 apresenta informações detalhadas sobre cada coleção.

Tabela 2. Descrição das bases de dados.

Base de Dados	#Doc.	Positivo	Negativo	Neutro	Termos Distintos
Séries de TV	3,117	27.94%	4.52%	67.53%	4,503
Eleições Americanas	1,353	20.33%	30.88%	48.79%	3,635

4.2. Configuração dos Experimentos

Como nosso objetivo é determinar o sentimento coletivo expresso em um conjunto de documentos, a avaliação de qualidade dos resultados é baseada na comparação da distribuição real de classes polares e a distribuição gerada por cada método. No caso do *SACI*, consideramos a distribuição de probabilidade das classes, e para os demais métodos consideramos a porcentagem de documentos individualmente assinalados a cada classe. Consideramos duas métricas tradicionais de aproximação de erro para comparar diferentes distribuições: (i) Distância Euclidiana (D.E.), que é a distância linear entre dois vetores e (ii) Distância Quadrática (D.Q.), que considera relações cruzadas entre pares de vetores. Ambas métricas assumem valores entre 0 a 1. Todos os resultados apresentados foram obtidos por meio de validação cruzada de *10-fold*, aplicando o *t-test* de calda dupla com 95% de confiança durante comparações feitas entre o *SACI* e os outros métodos.

Comparamos o *SACI* com quatro métodos de AS bem estabelecidos na área, sendo dois baseados em uma abordagem supervisionada: [Pang et al. 2012] (PLV) e [Dave et al. 2003] (DLP), e dois baseados em uma abordagem não-supervisionada: [Wilson et al. 2005] (WWH) e [Pak and Paroubek 2010] (PP), todos eles descritos na seção 2. Para os métodos PLV e DLP, utilizamos apenas unigramas, por apresentarem os melhores resultados, conforme discutido pelos autores. Ambas técnicas classificam documentos somente em positivo ou negativo. Para possibilitar a comparação entre todas as técnicas, adicionamos um passo que classifica documentos ente subjetivos e não subjetivos (neutros) e, posteriormente, os documentos classificados como subjetivos são classificados como positivo ou negativo[1].

[1]O código fonte de todos os algoritmos avaliados, bem como as coleções utilizadas, estão disponíveis em *http://dcomp.ufsj.edu.br/saciWeb/source.php*

4.3. Configuração de Parâmetros

Além do *lexicon* de entrada, o único parâmetro do *SACI* é o raio máximo de transformação l. Esse raio influencia diretamente a efetividade da análise e a escolha do valor apropriado depende do domínio de aplicação. Realizamos um experimento no qual o valor desse raio foi variado de 1 até o tamanho completo de cada sentença, e para cada valor de raio avaliamos a qualidade da análise feita pelo *SACI*. Observamos que valores pequenos geralmente não permitem que os termos que contêm algum sentimento relacionado ao nodo central (alvo da análise) seja considerado, por outro lado, valores grandes podem incluir termos que não necessariamente se referem ao tópico avaliado. Assim, raios de 3 a 6 tendem a selecionar um conjunto mais confiável de informações sobre cada tópico avaliado, em ambas as bases de dados utilizadas. Nos experimentos apresentados a seguir utilizamos um raio igual a 4 para ambas as bases.

4.4. Análise de Sentimento Coletivo pelo SACI

No intuito de avaliar a qualidade da análise de sentimento coletivo feita pelo *SACI*, realizamos um conjunto de experimentos para responder quatro questões principais relacionadas ao funcionamento do *SACI*.

1. *Como comparar uma análise coletiva a uma agregada?*

Comparamos análises agregadas derivadas de algoritmos tradicionais de AS com o *SACI* considerando métricas de erro aproximado em ambos os domínios. Agregamos a classificação individual de cada método tradicional, determinando a probabilidade de ocorrência de cada sentimento como sendo a porcentagem de documentos individuais atribuídos às classe. Na tabela 3 apresentamos os resultados relacionadas a essa comparação, onde o símbolo ▲ representa ganhos significativos, • representa ganhos ou perdas não significativos e ▼ perdas significantivas. Observamos que o *SACI* supera as outras técnicas na base Series de TV, a qual apresenta altos níveis de sobreposição, reduzindo o erro aproximado para 86.14% e 77.32%, nas métricas D.E. and D.Q., respectivamente. Por outro lado, o *SACI* mostrou ser estatisticamente comparável a análise agregada para a Eleições Americanas, uma vez que a sobreposição de informação nesse caso é reduzida devido a sua localidade temporal.

Tabela 3. Análise coletiva do *SACI* comparada com as outras técnicas.

Base	Series de TV				Eleições Americanas			
Métrica	D.E.		D.Q.		D.E.		D.Q.	
SACI	0.061		0.022		0.102		0.016	
	Valor	Ganho do SACI	Valor	Ganho do SACI	Valor	Ganho do SACI	Valor	Ganho do SACI
PLV	0,126	51,59% ▲	0,059	62,71% ▲	0,101	-0,99% •	0,014	-14,29% •
DLP	0,130	53,08% ▲	0,048	54,17% ▲	0,109	6,42% •	0,016	0,00% •
WWH	0,179	65,92% ▲	0,040	45,00% ▲	0,080	-27,50% •	0,014	-14,29% •
PP	0,440	86,14% ▲	0,097	77,32% ▲	0,104	1,92% •	0,026	38,46% ▲

2. *O SACI é bom para realizar classificações individuais?*

Apesar de não ser o foco do *SACI*, podemos adaptá-lo para realizar análises individuais de documentos. Nesse caso, avaliamos o sentimento de cada documento considerando apenas as sentenças relativas a ele e consideramos a classe de sentimento de maior probabilidade na análise individual de cada documento como a prevista pelo *SACI*. Comparamos as análises individuais realizas pelo *SACI* com as técnicas implementadas considerando tanto Acurácia quanto a métrica Macro-F1, conforme apresentado na tabela 4. Esses resultados mostram que uma boa classificação individual de sentimento não garante uma boa análise coletiva e vice-versa, corroborando com nossa hipótese de que uma identificação coletiva de opinião adequada

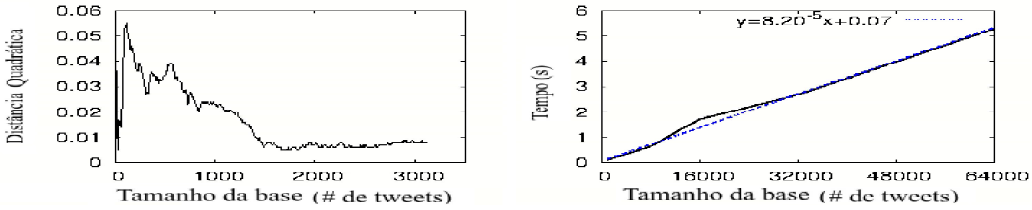
não requer uma análise previa e boa de opinião individual. Isso se dá pelo fato que a exploração da sobreposição de documentos traz novas informações úteis para AS.

Tabela 4. Comparação da análise individual do *SACI* com outras técnicas.

Base	Series de TV				Eleições Americanas			
Métrica	Acurácia (%)		Macro-F1 (%)		Acurácia (%)		Macro-F1 (%)	
SACI	79.25		61.30		52.11		47.96	
	Valor	Ganho do SACI	Valor	Ganho do SACI	Valor	Ganho do SACI	Valor	Ganho do SACI
PLV	78.18	1.37% ●	50.70	20.91% ▲	58.62	-11.11% ▼	54.22	-11.55% ▼
DLP	77.58	2.15% ●	65.84	-6.90% ▼	76.27	-31.68% ▼	74.08	-35.26% ▼
WWH	59.92	32.26% ▲	47.38	29.38 % ▲	45.73	13.95% ▲	42.02	14.14% ▲
PP	58.42	35.66% ▲	43.66	40.40 % ▲	64.59	-19.32% ▼	60.64	-20.91% ▼

3. *Quanto de informação o SACI precisa para realizar uma boa aproximação do sentimento coletivo real?*

Para responder essa questão, avaliamos o erro de aproximação variando o total de *tweets* analisados. Para esse experimento, utilizamos a base de Séries de TV, já que ela mantém uma boa localidade temporal dos tópicos discutidos. Consideraremos somente a métrica D.Q., já que ambas métricas apresentaram resultados similares, tal como mostrado na figura 3 (a). Quanto maior o número de tweets avaliados pelo *SACI*, menor é o erro. Além disso, 1500 tweets são suficientes para realizar uma aproximação bem próxima a melhor encontrada pelo *SACI*.



(a) Quantidade de Informação (b) Tempo de Execução.

Figura 3. Avaliação Quantitativa do SACI

4. *Qual é o custo computacional associado a análise coletiva do SACI?*

Eficiência é um requisito importante para AS em cenários de alta demanda. O *SACI* apresenta uma eficiente análise coletiva devido a três características principais: (1) o treino do *SACI* consiste na construção de um grafo sobre termos distintos observados em um conjunto de documentos; (2) o grafo é definido sobre o 'espaço de termos', que é muito menor que o 'espaço de postagens' disponíveis nas Mídias Sociais; e (3) com a chegada de novos dados o grafo pode ser atualizado, sem a necessidade de reconstrução do modelo. Para corroborar essa afirmação, realizamos uma análise de comportamento assintótica do *SACI*. Assumindo $|T|$ como o tamanho de todos os caminhos, a complexidade de analisar um único caminho é $O(|T|)$. No pior do caso, não há a sobreposição de caminhos. Para analisar todo o grafo com $|D| \times |F|$ caminhos distintos, onde $|D|$ é o número de documentos distintos a serem analisados e $|F|$ a média de sentenças distintas em cada documento, sua complexidade é $O(|D| \times |F| \times |T|)$. Como $|D|$ é maior que $|T|$ e $|F|$, a complexidade do *SACI* é linear quanto ao números de documentos distintos. Avaliamos o tempo de execução do *SACI* utilizando uma amostra maior das Séries de TV, em que as classes não são conhecidas, considerada apenas para análise de eficiência. Os resultados, considerando a média de 10 execuções em um *Intel Core i5* 2.4 GHz e 4 GB de RAM, são apresentados na figura 3 (b). De fato, a execução do *SACI* cresce de forma linear em relação ao número de tweets analisados.

Observamos que o *SACI* é capaz de processar aproximadamente 10000 *tweets* por segundo, evidenciando a aplicabilidade do *SACI* para análises em tempo real.

De forma a demonstrar tal aplicabilidade, consolidamos a partir do SACI uma ferramenta Web para análise de sentimento em tempo real sobre dados do Twitter. O uso dessa ferramenta para análise de tweets sobre o programa televisivo *BBB 14*[2], interessantemente, mostrou que além de ser capaz de prover análises em tempo real, o sentimento coletivo identificado coincidia com o resultado semanal do programa, apontando com alta precisão os candidatos mais rejeitados.

5. Conclusão e Trabalhos Futuros

Nesse trabalho apresentamos o *SACI*, um novo método de AS não-supervisionado baseado em *lexicon*, que define o sentimento coletivo de um conjunto de documentos a partir da avaliação do grafo de transição entre termos que incidem nesse conjunto. Demonstramos que uma análise mais precisa e eficiente de sentimentos coletivos pode ser alcançada por meio da classificação de conjuntos de documentos simultaneamente, explorando a sobreposição entre eles. Mais especificamente, reduzimos os erros de aproximação em até 86% em relação aos métodos tradicionais que agregam análise individual, com uma baixa complexidade computacional. Estes resultados apontam o *SACI* como um método AS coletivo promissor para fluxos de dados textuais curtos. De fato, a consolidação de uma ferramenta Web de AS em tempo real *SACI* evidencia sua utilidade prática. Como trabalho futuro, pretendemos adaptar o *SACI* para utilizar outros tipos de informações extraídas de mensagens, tais como os *emoticons* [Zhao et al. 2012].

Referências

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*.

Dave, K., Lawrence, S., and Pennock, D. M. (2003). Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *ACM WWW*.

Hu, X., Tang, L., Tang, J., and Liu, H. (2013). Exploiting social relations for sentiment analysis in microblogging. In *WSDM*, pages 537–546, Italy. ACM.

Mark, K. and Csaba, L. (2007.). Analyzing customer behavior model graph (cbmg) using markov chains. In *IEEE INES*, pages 71 –76.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*.

Pang, B., Lee, L., and Vaithyanathan, S. (2012). Thumbs up? Sentiment classification using machine learning techniques. In *EMNLP*, pages 79–86.

Sá, G., Silveira, T., Chaves, R., Teixeira, F., Mourão, F., and Rocha, L. (2014). Legi: Context-aware lexicon consolidation by graph inspection. In *ACM SAC*.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *EMLP*, pages 347–354, USA.

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

[2]Disponível em http://dcomp.ufsj.edu.br/saciWeb/bbb/