Revista de Informática Aplicada, Volume 11, Número 2, 2015

02

Análise de Opinião em Redes Sociais Envolvendo a Temática do Meio-Ambiente

Title: Opinion Mining in Social Networks Applied to Environment

André Luiz Firmino Alves[1]

Cláudio de Souza Baptista[2]

Roberta Falcão de Cerqueira Paes[3]

Anderson Almeida Firmino[4]

Luiz Henrique de Andrade[5]

Resumo: *A proliferação dos meios de comunicação na Web e a necessidade das empresas compreenderem os impactos de suas ações socioambientais junto à população afetada tornam necessária a adoção de um mecanismo automático para analisar as opiniões da população. Essa necessidade evidencia-se ainda mais no caso de empresas que tratam diretamente com alto risco ambiental. Nesse contexto, o objetivo deste trabalho é a aplicação de técnicas de análise de sentimentos em tweets relacionados à temática ambiental para auxiliar empresas do setor energético na análise dos impactos das ações empreendidas ao longo do tempo através das opiniões contidas nas mídias sociais. Neste trabalho, além de se utilizarem técnicas para classificação de sentimento nos tweets, um estudo utilizando os metadados contidos nos tweets (retweets, curtidas etc.) foi realizado com o propósito de identificar fatores não textuais que auxiliam na identificação de sentenças opinativas em tweets.*

Palavras-Chave: *Análise de Sentimentos. Twitter. Mineração de Opinião.*

Abstract: *The proliferation of social media on the Web and the need for companies to measure the impact of their environmental initiatives by the affected population has driven the proposal of an automatic mechanism to analyze the population's opinions. This necessity becomes even more evident especially regarding companies that deal directly with high environmental risk. In this context, the goal of this work is to apply sentiment analysis techniques on tweets related to environmental issues aiming to help energy companies analyze the impacts of environmental actions taken over time through opinions within social media. Besides using sentiment classification techniques in tweets, this work presents a study using tweet metadata (retweets , likes, ..) in order to identify non-textual features that help identifying opinionated sentences in tweets.*

Keywords: *Sentiment Analysis. Twitter. Opinion Mining.*

1. UFCG - Laboratório de Sistemas de Informação, Departamento de Sistemas e Computação, Paraíba, Brasil - andre@uepb.edu.br
2. UFCG - Laboratório de Sistemas de Informação, Departamento de Sistemas e Computação, Paraíba, Brasil - baptista@dsc.ufcg.edu.br
3. Companhia Hidro Elétrica do São Francisco, Pernambuco, Brasil - rfcpaes@chesf.gov.br
4. UFCG - Laboratório de Sistemas de Informação, Departamento de Sistemas e Computação, Paraíba, Brasil - andersonalmeida@copin.ufcg.edu.br
5. UFCG - Laboratório de Sistemas de Informação, Departamento de Sistemas e Computação, Paraíba, Brasil - luizha.cc@gmail.com

1 Introdução

Nas últimas décadas, nota-se um crescimento na preocupação ambiental pelos diversos atores sociais - governantes, empresas e cidadãos - no tocante aos impactos ambientais decorrentes da ação humana. Tal preocupação tem resultado numa conscientização da sociedade, colocando-a como fiscal social com relação à degradação ambiental, e demandando leis ambientais cada vez mais rigorosas (NICOLELLA; MARQUES; SKORUPA, 2004). Particularmente, no setor elétrico brasileiro, cuja base energética está nas hidroelétricas, que trazem grandes danos ambientais decorrentes das inundações geradas quando de suas implantações, há uma crescente preocupação de como mitigar esses impactos por meio de ações socioambientais junto à população afetada. Nesse sentido, a política ambiental atual para o setor elétrico aduz que a gestão ambiental deve englobar do planejamento à implantação e manutenção dos empreendimentos energéticos.

Resultados em direção à melhoria no tratamento aos impactos ambientais causados por empresas de energia passam pela adoção de Sistemas de Gestão Ambiental (SGA), que visam implantar políticas e procedimentos técnico-administrativos, possibilitando um melhor controle dos impactos ambientais causados. Nesse processo de gestão ambiental, há uma necessidade de envolver a sociedade civil com o objetivo de implantar uma nova maneira de tratar as questões ambientais, promovendo, assim, a construção da cidadania e do fiscal social. Assim, compreender o que as pessoas estão pensando ou suas opiniões é fundamental para tomada de decisão, especialmente no contexto em que as pessoas expressam seus comentários de forma voluntária, principalmente através das diversas mídias sociais (blogs, fóruns de discussões, microblogging, redes sociais etc.), proporcionados pelo surgimento da Web 2.0.

Analisar os comentários expressos nas mídias sociais é uma tarefa difícil de ser realizada manualmente, principalmente devido ao grande volume de dados. Nesse contexto, a técnica de análise de sentimentos pode ser aplicada para compreender as opiniões da sociedade sobre a questão ambiental propagada em mídias sociais. De acordo com Liu (2012), a análise de sentimentos tem como principal objetivo obter e formalizar a opinião e o conhecimento subjetivo em documentos não estruturados (textos), para posterior análise dentro de um domínio específico. Visa, pois, analisar as opiniões, sentimentos, avaliações, atitudes e emoções de pessoas com relação a entidades, pessoas, produtos, serviços, eventos, dentre outros. Uma forma comum de compreender o sentimento geral de um conjunto de documentos é através da sumarização das opiniões realizada através das classificações das opiniões com base em categorias (Polaridade): positiva, negativa e neutra (LIU, 2012; PANG; LEE, 2008).

Nos últimos anos, a ferramenta Twitter tem sido largamente utilizada para expressar opiniões diversas, tornando-se um repositório bastante atrativo para se implantar técnicas de análise de sentimentos (KWAK et al., 2010). Todavia, devido à informalidade textual expressa pelos usuários, a tarefa de análise de sentimentos em tweets torna-se por demais complexa. Ademais, esses problemas são agravados pela limitação de 140 caracteres e pela complexidade gramatical da língua portuguesa.

Por outro lado, na maioria das pesquisas na área de análise de sentimentos apenas a informação textual de cada tweet é analisada. Segundo Suh et. al. (2010), um tweet contém, além da informação textual, propriedades de conteúdo e contextuais. As propriedades de conteúdo que podem ser encontradas no tweet incluem URL's, hashtags e menções (referências a outros usuários). Já as propriedades contextuais incluem número de seguidores do usuário, número de curtidas do tweet, número de retweets, dentre outras. De acordo com Harris et. al. (2015), curtir um tweet demonstra concordância do usuário com o conteúdo do tweet ou com a opinião dele. Dessa forma, se houver um tweet com polaridade de sentimento positiva e 10 curtidas, indica que, além do autor do tweet, outras 10 pessoas concordam com a opinião dele (MEIER; ELSWEILER; WILSON, 2014). Assim, ao explorar características adicionais de tweets é possível obter uma análise de sentimento mais completa, contemplando também o impacto que esse tweet tem em relação a seus seguidores.

Portanto, o principal objetivo deste trabalho é aplicar técnicas de análise de sentimentos em tweets escritos em português para auxiliar na atividade de compreensão do sentimento da sociedade relacionada à temática ambiental. Algoritmos de aprendizagem de máquina são utilizados para detecção da polaridade do sentimento. Para a sumarização do sentimento foram utilizados gráficos com variação temporal da polaridade

detectada. A ferramenta desenvolvida com este estudo é aplicável dentro de um SGA - Sistema de Gestão Ambiental - para possibilitar uma observação detalhada de alto nível e global, auxiliando, dessa forma, na tomada de decisões.

Este trabalho é uma extensão de Alves et al. (2015), apresentando como principal contribuição um estudo sobre os atributos presentes nos tweets que contribuem com a identificação de sentenças opinativas, visando melhorar a tarefa de classificação de polaridade. Os atributos de interesse deste trabalho são curtidas, menções a usuários, retweets e links nos tweets. Neste estudo é realizado uma investigação com base em análise estatística para verificar se existem atributos significativamente importante para identificação de conteúdo opinativo em tweets e quais seriam esses atributos.

O restante deste artigo está organizado como segue. Na seção 2, são apresentados alguns trabalhos relacionados. Na seção 3, apresenta-se a metodologia considerada para o desenvolvimento do analisador de sentimentos. Na seção 4, apresentam-se os resultados obtidos. Por fim, na seção 5 são apresentadas as conclusões e os trabalhos futuros.

2 Trabalhos Relacionados

A análise de sentimentos tem sido uma das áreas de pesquisa mais ativas no campo de *Natural Language Processing* - NLP (PANG; LEE, 2008). Uma visão geral sobre análise de sentimentos pode ser encontrada em Pang e Lee (2008). Diversos domínios têm aplicado técnicas de análise de sentimentos tais como: mercado acionista, através da identificação do humor do mercado baseado nas opiniões de especialistas (O'HARE et al., 2009); opiniões dos consumidores acerca de produtos ou serviços (EIRINAKI; PISAL; SINGH, 2012; HU; LIU, 2004); aplicações de turismo, através da análise dos comentários dos viajantes (BJØRKELUND; BURNETT; NØRVK, 2012) análise de políticos e de política (FANG et al., 2012).

As principais técnicas para classificar a polaridade do sentimento em documentos não estruturados têm sido baseadas em aprendizado de máquina, análise semântica, estatística e análise léxica. No entanto, técnicas de aprendizagem de máquina têm sido mais exploradas na literatura (SHARMA; DEY, 2012; FELDMAN, 2013). Uma das principais limitações no uso de aprendizado supervisionado é a necessidade de dados rotulados para treinamento e teste (*dataset*). Para auxiliar a atividade de coletar os dados rotulados de forma automática,vários trabalhos fizeram uso de *emoticons* – caracteres que transmitem emoções. Em Li e Li (2011) verifica-se que 87% dos *tweets* contendo *emoticons* possuem os mesmos sentimentos representados pelos *emoticons* no texto. Trabalhos que utilizam *emoticons* para treinamento dos classificadores têm apresentado excelentes resultados de acurácia (acima de 80%) (PAK; PAROUBEK, 2010).

Por outro lado, ainda existem poucos trabalhos na literatura que realizam análise de sentimentos utilizando um *corpus* em língua portuguesa. Os trabalhos de Chaves et al. (2012), Sarmento et al. (2009) e Tumitan e Becker (2013) utilizam técnicas de análise léxica baseada em dicionários e apenas o trabalho de Nascimento et al. (2012) utiliza técnicas de aprendizagem de máquina.

Em Chaves et al. (2012) é apresentado um algoritmo que utiliza uma abordagem de análise léxica para classificação de sentimentos em comentários em português. Ontologias e lista de adjetivos polarizados (positivo, negativo e neutro) que expressam sentimentos são utilizados para definir a orientação semântica dos textos analisados. Os resultados indicam uma média do F-Measure de apenas 0,32 no reconhecimento da polaridade. Tumitan e Becker (2013) utilizam um dicionário de palavras (SentiLex-PT), que foi adaptado segundo o contexto, para analisar as opiniões em comentários sobre políticos realizados em jornais e estudar a correlação dos sentimentos expressos com as pesquisas de intenção de votos. Em Nascimento et al. (2012), classificadores de sentimento são utilizados para avaliar as reações das pessoas no Twitter em relação às notícias vinculadas na mídia. Os resultados de acurácia variaram de 70% a 80% de acordo com o tipo de notícia e classificador utilizado.

Outros estudos iniciam a análise das propriedades contextuais de um *tweet*, como, por exemplo, o número de *retweets*. Meier, Elsweiler e Wilson (2014) realizaram um estudo para entender o comportamento da funcionalidade "curtir" do Twitter. Segundo eles, "*retweetar*" indica que o usuário considera a informação suficientemente interessante para ser retransmitida para seus seguidores. Já o "curtir" indica que o usuário apenas concorda com o conteúdo do *tweet*.

Pesquisas na literatura tentaram estabelecer alguma relação entre algumas das propriedades contextuais de conteúdo do *tweet* e sua opinião. Stieglitz e Dang-Xuan (2012) e Pfitzner, Garas e Schweitzer (2012) estabeleceram uma relação entre a opinião de um *tweet* e sua retweetabilidade. Segundo Stieglitz e Dang-Xuan (2012), *tweets* que contém mais palavras com sentimento positivo ou negativo tendem a ser mais retweetados. Já segundo Pfitzner, Garas e Schweitzer (2012), *tweets* emocionalmente diversificados, isto é, contendo palavras com sentimentos positivos e negativos, têm até cinco vezes mais chances de serem retweetados.

A abordagem para análise de sentimento proposta neste trabalho difere dos trabalhos relacionados aqui analisados no tocante ao uso de dois classificadores, evitando o uso de PosTagger (Part-of-Speach) na identificação de um conteúdo opinativo. Destarte, o primeiro classificador identifica se um conteúdo é subjetivo ou objetivo; e o segundo classificador identifica a polaridade (positiva ou negativa) do conteúdo já identificado como subjetivo (opinativo). Outro diferencial deste trabalho foi o levantamento de quais características são mais relevantes para a análise de sentimentos no contexto da rede social Twitter.

3 Processo de Análise de Sentimentos

Nesta seção são abordadas a metodologia utilizada no processo de análise de sentimentos bem como a formalização do problema a ser endereçado.

3.1 Visão Geral da Metodologia

A abordagem de análise de sentimento proposta neste trabalho é caracterizada pela utilização de técnicas de Recuperação da Informação (RI) para prover suporte à tomada de decisão através da mineração de informações contidas em microtextos, como os de microblogging e de redes sociais. Especificamente, este trabalho utiliza técnicas de análise de sentimentos para determinar a polaridade da opinião (positiva ou negativa) utilizando textos oriundos do microbloging Twitter. A Figura 1 apresenta a visão geral da metodologia utilizada no processo de análise de sentimento abordado neste trabalho.

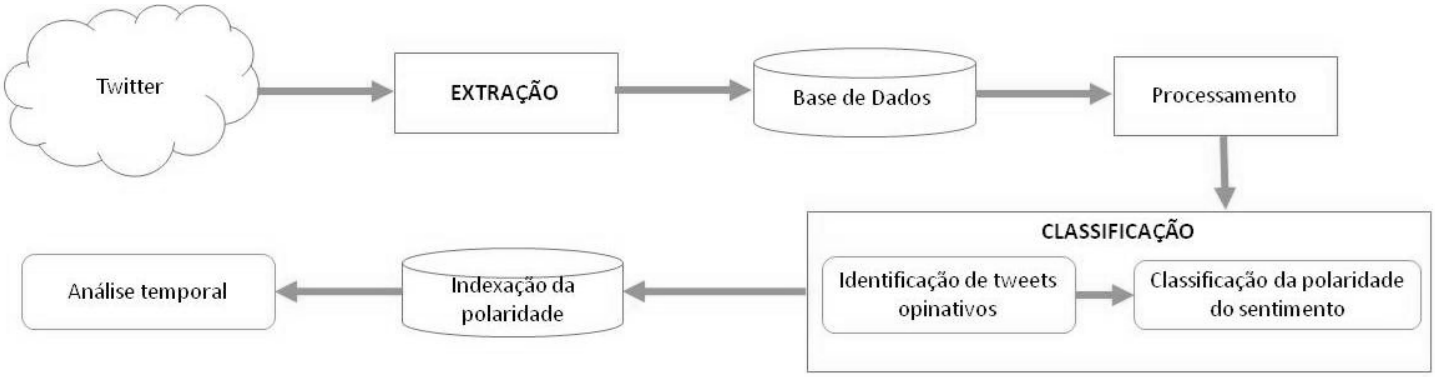


Figura 1 - Visão Geral do processo de Análise de Sentimento
Fonte: dados da pesquisa.

O primeiro passo é a coleta dos dados a partir do Twitter. No segundo passo, os tweets são submetidos à uma fase de pré-processamento, que inclui a remoção de stopwords (preposições, artigos etc.), termos especiais usados no Twitter (RT, via etc.), remoção de nomes dos usuários e tratamento de hashtags (#) (separação de termos compostos de acordo com a capitalização das letras). No terceiro passo, contempla-se a etapa da classificação de polaridade de sentimento, que se constitui no principal componente e objeto de estudo de várias pesquisas da área de análise de sentimentos, cujo objetivo é a identificação dos sentimentos contidos nos textos. Neste trabalho, a partir dos resultados apresentados por Alves et al. (2014) em estudos comparativos entre técnicas de classificação da polaridade, a definição da polaridade é realizada utilizando dois modelos de classificação: o primeiro classifica os tweets em opinativos (subjetivos) ou informativo (objetivo) e o segundo classifica a polaridade (Positivo ou Negativo) dos tweets opinativos. Finalmente, após a indexação dos sentimentos detectados pela etapa de classificação, atinge-se a última etapa da metodologia, na qual os resultados são explorados através de uma análise temporal dos sentimentos, obtendo, assim, a sumarização da análise de sentimentos.

3.2 Formalização do problema

Os microtextos explorados nesta pesquisa foram oriundos do *microblogging* Twitter, embora a proposta desta pesquisa seja aplicável a outros tipos de microtextos, como os de redes sociais. Ademais, consideramos que uma mensagem do Twitter apresenta apenas um sentimento predominante, embora algumas pesquisas considerem múltiplas emoções associadas a uma mensagem.

Seja um *tweet* $t_i \in T$, tem-se que a principal atividade da classificação do sentimento é obter a opinião predominante sobre as expressões textuais do documento t_i. O problema de detecção de polaridade do sentimento pode ser tratado como uma tarefa de categorização de textos. Mais formalmente, uma tarefa de classificação é encontrar uma função que aproxima a uma função de classificação $f: T \rightarrow C$, com $f(t_i) = c_j$, tal que $C = \{c_1,..c_n\}$ representa um conjunto de n categorias predefinidas. A função f descreve como os textos são associados às classes, atribuindo um texto $t_i \in T$ para sua categoria $c_j \in C$. Neste trabalho, o conjunto T representa todos os textos coletados e $c_j \in C = \{positivo, negativo\}$ é a polaridade predominante (orientação semântica do sentimento) relacionada ao *tweet* t_i.

3.3 Identificação da polaridade do sentimento

Para definir a função de classificação $f: T \rightarrow C$ para a identificação do sentimento nos *tweets* foi utilizada uma abordagem de aprendizagem supervisionada de máquina, que necessita de um conjunto de dados (*dataset*) já rotulados formado por dois conjuntos disjuntos: treinamento e teste (validação). Então, dados dois conjuntos T_t e T_v, respectivamente de treinamento e validação, tem-se que $T_t \cap T_v = \theta$. Assim, a classificação supervisionada começa com o conjunto de treinamento ($T_t = t_1,..,t_n$) com os textos já marcados com as categorias $c_j \in C = (c_1,..,c_n)$ e a tarefa é determinar o modelo de classificação capaz de correlacionar corretamente um novo texto $t_w \in T_v$ à sua categoria, ou seja, $f: T \rightarrow C$ com $f(t_w) = c_j$. Neste caso, o conjunto de treinamento é utilizado para a construção de um classificador que aprenderá automaticamente as regras e características gerais dos documentos classificados. Já o conjunto de teste faz-se necessário para validar o classificador treinado.

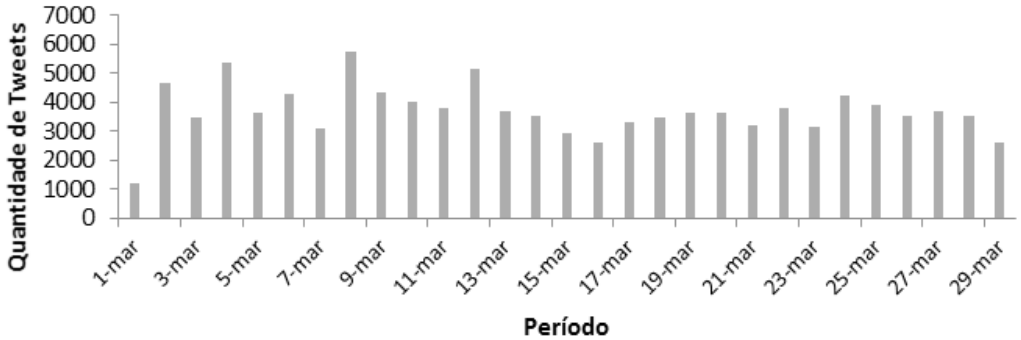


Figura 2 - Quantidade de **tweets** coletados durante o período considerado
Fonte: dados da pesquisa.

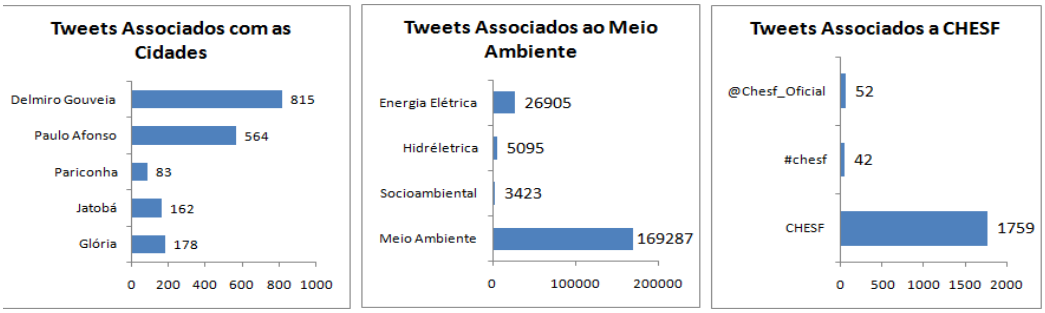


Figura 3 - Quantidade de **tweets** associados com os termos da coleta
Fonte: dados da pesquisa.

Na tarefa de classificação, devem-se considerar as características (*features*) que o modelo de classificação observa nos dados de treinamento no processo de aprendizagem. Na classificação de textos, as porções (*tokens*) dos textos são extraídas e analisadas, e o classificador seleciona as características relevantes, representando-as na forma de um vetor de termos (*bag of words*). Neste trabalho, apenas a característica de frequência de termos foi utilizada nos classificadores implementados.

4 Estudo de Caso

Nesta seção é apresentado um estudo de caso junto à empresa do setor elétrico Companhia Hidroelétrica do São Francisco - CHESF, que visa validar a metodologia aqui proposta.

4.1 Seleção do *Corpus*

O objetivo desse trabalho é analisar as opiniões da sociedade relacionadas à temática ambiental. Dessa forma, tomou-se como estudo de caso o Plano de Ação Socioambiental (PAS), idealizado pela Chesf. O PAS é um conjunto de ações que envolve toda a sociedade civil com o objetivo de implantar uma nova maneira de tratar as questões ambientais, promovendo, assim, a construção da cidadania. O PAS abrange a área que fica circunscrita à área de influência do complexo hidrelétrico de Paulo Afonso, constituída por cinco municípios: Paulo Afonso - BA, Glória – BA, Delmiro Gouveia - AL, Pariconha – AL e Jatobá-PE.

O Twitter foi utilizado como fonte de dados para realização da análise de sentimentos. Para coletar os dados, foi criado um motor de busca, que utiliza a API (*Application Programming Interface*) do Twitter e, diariamente, durante o período de coleta, armazenava *tweets* que continham termos previamente selecionados. A Figura 2 apresenta o quantitativo de *tweets* coletados durante o mês de Março de 2015. No total, foram coletados cerca de 107.000 *tweets*.

Os termos utilizados na coleta dos dados foram definidos segundo três parâmetros: 1) cidades de atuação do PAS; 2) Termos relacionados ao meio ambiente; e 3) Termos relacionados à CHESF. Os gráficos da Figura 3 apresentam a distribuição do quantitativos de *tweets* associados com os termos utilizados na coleta segundo esses três critérios mencionados.

Tabela 1 - Exemplo de **Tweets** Rotulados Manualmente

Rótulo	***Tweet***
Positivo	"a parte da hidreletrica foi uma das melhores"
Negativo	"É impressão minha ou a energia elétrica está tendo aumento de 15 em 15 dias? Isso é uma vergonha!"
Neutro	"A competência para legislar sobre meio ambiente do trabalho não é concorrente mas privativa da União. STF ADI 1893-MC"

Fonte: dados da pesquisa.

Tabela 2 - Quantidade de **Tweets** Rotulados (Conjunto de Treinamento e Testes)

	Quantidade			
Abordagem	**Positivos**	**Negativos**	**Neutro**	**Total**
Emoticons	126	73	-	199
Rotulação Manual	181	188	1409	1778

Fonte: dados da pesquisa.

4.2 Métricas para Avaliação

As métricas utilizadas para avaliação dos resultados dos algoritmos de detecção de polaridade foram as utilizadas frequentemente na literatura em avaliações de sistemas de Recuperação da Informação (RI): acurácia, precisão, revocação e F-measure. Especificamente, na área de análise de sentimentos, são as métricas utilizadas na literatura para avaliação dos algoritmos de detecção de polaridade dos sentimentos.

Para a avaliação do algoritmo que utiliza técnicas de aprendizado de máquina supervisionado para identificação da polaridade do sentimento, uma fração dos *tweets* rotulados é reservada para treinar o modelo, não sendo utilizada para a obtenção das métricas. A outra fração é utilizada para aplicar o classificador de sentimento e comparar os resultados com os rótulos marcados. Dessa forma, adotamos duas abordagens para obtenção de *tweets* com sentimentos já rotulados:

- *Emoticons*: utilizou a abordagem de Pak e Paroubek (2010) na qual considera-se que toda as palavras contidas na mensagem que contém os caracteres que expressam emoções, e.g. *emoticon* alegre - ":-)", ":)", "=)", ":D", etc - e *emoticon* triste ":-(", ":(", "=(", ";(" etc -, também estão associadas à emoção do caractere (*emoticon*). Assim, se um *tweet* apresenta um *emoticon* alegre

(":-)"), por exemplo, sua polaridade considerada é positiva.

• Rotulação Manual: 1.778 *tweets* dentre os coletados foram escolhidos de forma aleatória e separados para rotulação manual da polaridade do sentimento.

A Tabela 1 apresenta exemplos de twitters que foram rotulados manualmente. Os dois métodos de obtenção de rotulação de sentimentos foram utilizados na comparação e combinação de resultados dos classificadores. A Tabela 2 apresenta o quantitativo de *tweets* com sentimentos rotulados a partir de cada abordagem.

Para avaliar a capacidade de generalização dos modelos de classificação, foi utilizado o método de validação cruzada *k-fold*, com k=10, ou simplesmente *10-fold*.

4.3 Classificadores Utilizados na Definição da Polaridade

O processo de classificação da polaridade é realizado com a utilização de dois classificadores. O primeiro classificador é utilizado para classificar os textos em objetivos e subjetivos, funcionando como uma espécie de filtro para o próximo classificador. Já o segundo classificador foi treinado para classificar as polaridades das mensagens subjetivas em apenas duas classes: Positiva e Negativa.

Inicialmente, no planejamento do experimento deste estudo de caso, foi cogitada a utilização de classificadores SVM (*Suport Vector Machine*), considerando que esse classificador tem apresentado bons resultados em problemas de classificação (SHARMA; DEY, 2012; PANG; LEE, 2008; READ, 2005). No entanto, considerando o conjunto de *tweets* rotulados (*dataset*), conforme Tabela 2, observamos um desbalanceamento entre as classes dos *tweets* rotulados. Isso se deve pelo fato de que, no contexto aplicado, o número de notícias relacionadas ao tópico é notoriamente maior que o número de opiniões relatadas. Assim, ao escolher o classificador utilizado neste trabalho, considerou-se o desempenho do algoritmo com uma base de dados desbalanceada (AKBANI; KWEK; JAPKOWICZ, 2004). Portanto, os classificadores Naive Bayes foram utilizados neste trabalho, uma vez que se aplicam perfeitamente com a premissa de dados independentes e com classes não balanceadas.

Tabela 3 - Resultados do Classificador de Sentimentos

Classificador	Acurácia	Precisão	Revocação	F-Measure
Naive Bayes-Classificação Dupla	0,721	0,730	0,721	0,725

Fonte: dados da pesquisa.

4.4 Características de *tweets* que podem auxiliar na identificação de conteúdo opinativo

A identificação de um conteúdo opinativo é uma tarefa muito importante na área de análise de sentimento. A metodologia implementada em outros trabalhos de análise de sentimentos utiliza apenas as informações textuais dos *tweets* (máximo de 140 caracteres) (LI; LI, 2011; PAK; PAROUBEK, 2010; TUMITAN; BECKER, 2013). E uma análise de outros elementos não textuais pode ser muito importante da identificação de *tweets* opinativos.

No entanto, um *tweet* contém, além do texto da mensagem escrito pelo autor, outras informações que são preenchidas implicitamente pelo Twitter. Esses metadados podem documentar, por exemplo, a hora e a localização geográfica do usuário no momento do envio. Ignorando por ora os textos dos *tweets*, e objetivando verificar quais das propriedades de um *tweet* podem ser utilizadas na detecção de *tweets* opinativos, este trabalho explorou os seguintes metadados dos *tweets*:

(i) curtidas (favoritos) - Indica se o *tweet* foi marcado como favorito (curtido) por algum usuário;

(ii) *retweets* - Indica se o *tweet* foi motivo de outro *tweet* realizado por outro usuário;

(iii) menções a outros usuários - Quantifica menções a outros usuários da rede;

(iv) *links* (urls) no texto - Indica se o *tweet* contém *links* para *sites* externos.

Desse modo, este estudo visa-se auxiliar na tarefa de identificação de *tweets* opinativos, através de uma análise de correlação entre os metadados dos *tweets* elencados e o fato de um *tweet* ser opinativo.

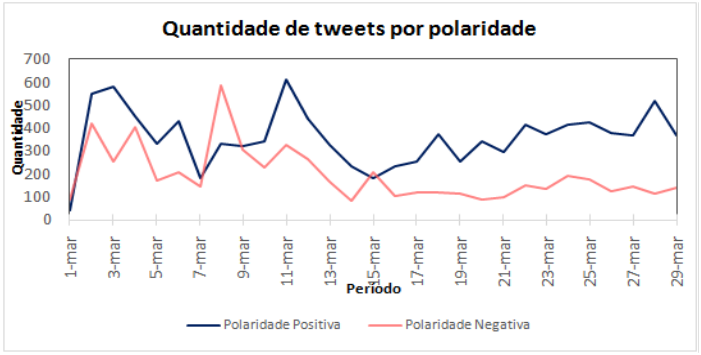
4.5 Resultados

4.5.1 Classificação da polaridade

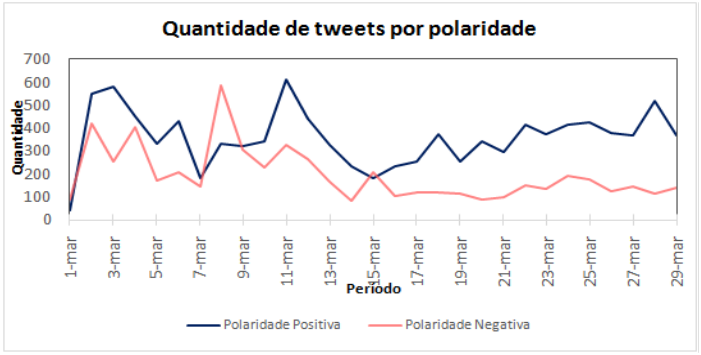
A Tabela 3 apresenta, através do processo de validação cruzada 10-fold, os resultados (média ponderada) obtidos pelo processo de classificação da polaridade. Os resultados são considerados excelentes para *tweets* no idioma português, principalmente pelo fato do desbalanceamento apresentado na base de dados rotuladas (*dataset*), onde a fração de *tweets* objetivos (polaridade neutra) é muito maior quando comparado com os *tweets* com polaridades positivas e negativas.

Aplicando esse classificador a todos os *tweets* coletados, as polaridades foram definidas e indexadas para análise na etapa de sumarização do sentimento. Do total de *tweets*, apenas 16.134 (cerca de 15%) foram considerados pelo classificador como *tweets* que apresentam sentimentos (positivos ou negativos). Os gráficos da Figura 4 ilustram o comportamento do sentimento detectado ao longo do período. Na Figura 4.a ilustra a distribuição temporal contendo o comportamento da quantidade de sentimentos positivos e negativos por dia. Esse gráfico é utilizado quando deseja-se conhecer de forma global, considerando inclusive o fator quantitativo na análise, os sentimentos expressos pelos usuários ao longo do período analisado. Já a Figura 4.b ilustra as proporções de *tweets* positivos e negativos considerando apenas as quantidade de *tweets* por dia. Por exemplo, percebe-se que no dia 28 de Março, cerca de 80% de *tweets* do dia apresentam sentimentos positivos e apenas 20% com sentimentos negativos.

Uma forma de obter a orientação semântica geral do sentimento expresso nos microtextos é através da subtração do número de *tweets* com sentimentos positivos pela quantidade de *tweets* com sentimento negativo. A Figura 5 apresenta um gráfico que ilustra a orientação semântica. O objetivo desse gráfico é ilustrar a tendência do sentimento em relação ao tema analisado. Por exemplo, percebe-se que no período de 4 a 7 de Março há uma mudança de sentimento e, dessa forma, há possibilidade de analisar os motivos dessa variação de sentimentos. No gráfico, também utilizamos a média móvel para suavizar a curva, eliminando os efeitos de valores atípicos (*outliers*).



(a)Quantidade de *tweets* por polaridade



(b)Fração de *tweets* por polaridade

Figura 4 - Comportamento da polaridade detectado sobre todos os **tweets** coletados
Fonte: dados da pesquisa

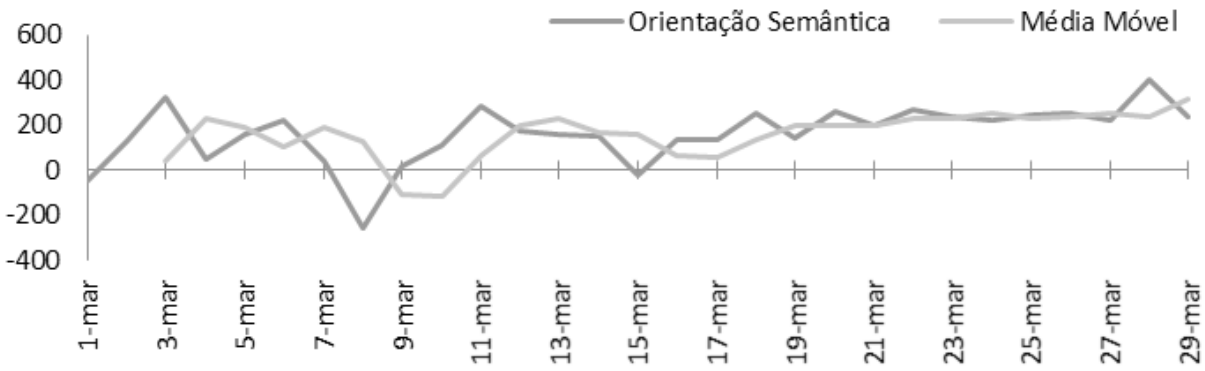


Figura 5 - Orientação Semântica Geral
Fonte: dados da pesquisa

4.5.2 Identificação dos atributos relevantes em sentenças opinativas

Para gerar o modelo de regressão logística optou-se por utilizar o *gold dataset* em detrimento do conjunto dos *tweets* rotulados pelo classificador a fim de minimizar a introdução de erros no modelo. Assim, o conjunto de todos os *tweets* coletados (cerca de 107.000) foi utilizado apenas para realizar uma comparação da disposição dos metadados nesses *tweets* com os demais. Os *tweets* que continham a polaridade do sentimento classificada como positiva ou negativa foram consideradas como *tweets* opinativos, e os classificados com polaridade neutra foram considerados como *tweets* informativos.

A Figura 6 apresenta o sumário de distribuição dos *tweets* (tanto do conjunto total quanto *do gold dataset*) de acordo com os metadados analisados. Pode-se perceber que em ambos os conjuntos existem mais *tweets* sem curtidas, sem menções a outros usuários e sem *retweets*. No *gold dataset*, vê-se que há mais *tweets* contendo *links*, enquanto que no conjunto total há mais *tweets* sem *links*. No geral, pode-se verificar que o *gold dataset* é representativo em relação ao conjunto total de *tweets.*

Um modelo de regressão logístico foi gerado para verificar quais atributos de um *tweet* seriam significativos para a identificação de sentenças opinativas. Mediante análises estatísticas, foi verificado que os atributos *link*, curtida e *retweet* são significativos em se tratando do fato de um *tweet* ser opinativo.

Ao analisar os dados utilizados no experimento, pode-se visualizar o impacto dos metadados estudados na opinião de um *tweet*. Como pode-se ver na Figura 7, os metadados que mais contém *tweets* opinativos são *links*, *retweets* e curtidas, reforçando, assim, que o fato de um *tweet* conter algum desses metadados está relacionado com o fato de um *tweet* ser opinativo.

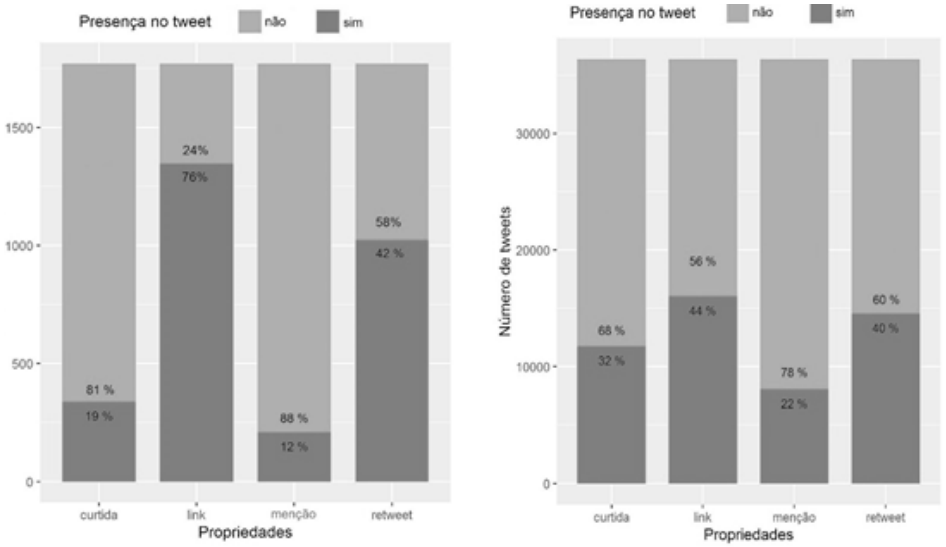


Figura 6 - Percentual de presença de propriedades nos **tweets** (**gold dataset** à esquerda e todos os **tweets** à direita)
Fonte: dados da pesquisa.

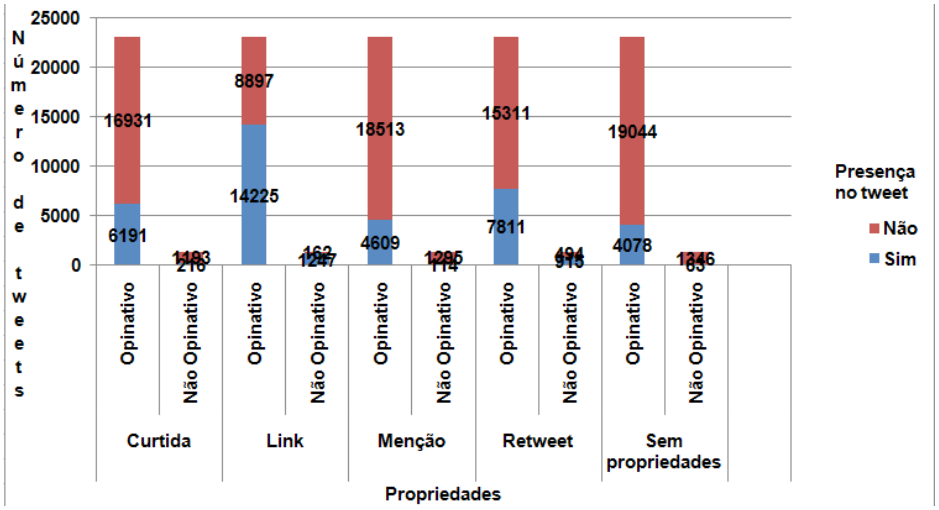


Figura 7 - Visão Geral do processo de Análise de Sentimento
Fonte: dados da pesquisa.

5 Considerações Finais

Neste artigo, foi apresentado um estudo de caso que utiliza técnicas de análise de sentimentos aplicadas a *tweets* relacionados ao meio-ambiente. Um modelo de classificação Naive-Bayes foi treinado e validado, obtendo excelentes resultados para *tweets* escritos no idioma português. Mesmo o conjunto de *tweets* coletados apresentando majoritariamente *tweets* objetivos, aproximadamente 85% dos *tweets*, foi possível definir um classificador que identifica a polaridade do sentimento de um *tweet* com uma acurácia de 72,1% e F-Measure de 72,5%.

Identificado os *tweets* opinativos de todo conjunto coletado e realizando classificação da polaridade dos sentimentos, através de técnicas de aprendizagem de máquina, foi possível sumarizar as opiniões através de uma análise temporal sobre os dados, possibilitando a identificação da orientação semântica dos sentimentos expressos pelos usuários. Embora a técnica de sumarização da opinião tenha sido aplicada a todo o conjunto de dados coletado, através dos termos das pesquisas, é possível realizar a sumarização individualmente sobre cada termo de interesse. Tomando por exemplo este estudo de caso, é possível realizar uma análise de sentimentos nos *tweets* das cidades de atuação do PAS, analisando o impacto ou a correlação das ações ambientais promovidas pela CHESF nos sentimentos expressados pela sociedade nas redes sociais ou microbloggings. Além do mais, utilizando técnicas avançadas para identificar o objeto avaliado em um *tweet* opinativo, a exemplo de Reconhecimento de Entidades Nomeadas (NER), é possível identificar o sentimento da sociedade em relação à empresa CHESF.

Outra contribuição deste trabalho consiste em um modelo de regressão logístico, o qual resultou nas seguintes conclusões:

O fato de um *tweet* conter menções a outros usuários não é determinante para dizer se ele é opinativo. Isso se deve ao fato de que muitos dos *tweets* que não são opinativos (ex.: notícias) também contêm menções.

Um *tweet* conter *links* para *sites* externos, ter *retweets* ou ter curtidas parece ser determinante para dizer se ele é opinativo. Isso se deve ao fato de inúmeros *tweets* tecerem comentários a respeito de alguma notícia de outros *sites*.

A abordagem desenvolvida neste estudo é aplicável dentro de um SGA para possibilitar uma observação detalhada de alto nível e global, auxiliando, dessa forma, na tomada de decisões. Um trabalho futuro é aplicar essas técnicas em um ambiente corporativo, no qual são coletadas as opiniões dos agentes envolvidos no PAS e, assim, possibilitará compreensão das opiniões de forma automática desses agentes em relação às ações empreendidas ao longo do tempo. Além do mais, através do estudo das características de um *tweet*, a identificação de propriedades relevantes para identificação de um *tweet* opinativo será incorporada ao classificador de sentimento, adicionando tais características ao modelo do classificador, com o objetivo de melhorar os resultados do processo de classificação de polaridade.

Agradecimentos

Os autores agradecem o suporte financeiro da ANEEL, sob o contrato de P&D+I Nº ANEEL 0048-1119/2012.

Referências

AKBANI, R.; KWEK, S.; JAPKOWICZ, N. Applying support vector machines to imbalanced datasets. In: EUROPEAN CONFERENCE ON MACHINE LEARNING - ECML, 15., 2004, Italy. Proceedings... Italy, 2004. p. 39-50.

ALVES, A. L. F.; BAPTISTA, C. S.; ANDRADE, L.H; PAES, R.F.C. Uso de Técnicas de Análise de Sentimentos em Tweets relacionados ao Meio-Ambiente. In: **WORKSHOP DE COMPUTAÇÃO APLICADA À GESTÃO DO MEIO AMBIENTE E RECURSOS NATURAIS -** WCAMA, 5., 2014, Brasília. Anais... Brasília: BDBCOMP, 2015.

Alves, A. L. F., Baptista, C. S., Firmino, A. A., Oliveira, M. G., Figueirêdo,H. F.. Temporal Analysis of Sentiment in Tweets: A Case Study with FIFA Confederations Cup in Brazil. DEXA, v.1, p. 81-88, 2014.

BJØRKELUND, E.; BURNETT, T. H.; NØRVK. A study of opinion mining and visualization of hotel reviews. In: International Conference on IIWAS, 14., 2012, New York-USA. Proceedings… New York-USA: ACM Press, 2012. p. 229.

CHAVES, M.; DE FREITAS, L.; SOUZA, M.; VIEIRA, R. PIRPO: An Algorithm to Deal with Polarity in Portuguese Online Reviews from the Accommodation Sector. Natural Language Processing and Information Systems 7337, p. 1–5, 2012.

EIRINAKI, M.; PISAL, S.; SINGH, J. Feature-based opinion mining and ranking. Journal of Computer and System Sciences, n.78, v.4, p. 1175–1184, July 2012.

FANG, Y.; SI, L.; SOMASUNDARAM, N.; YU, Z. Mining contrastive opinions on political texts using cross-perspective topic model. In: ACM INTERNATIONAL CONFERENCE ON WEB SEARCH AND DATA MINING – WSDM, 5., 2012, New York. Proceedings… New York-USA: ACM Press, 2012. p. 63.

FELDMAN, R. Techniques and applications for sentiment analysis. Communications of the ACM, n. 56, v.4, p. 82, Apr. 2013.

HARRIS, J.K.; MART, A.; MORELAND-RUSSELL, S.; CABURNA,Y. C. Diabetes Topics Associated With Engagement on Twitter. Prev Chronic Dis, v. 12, 2015.

HU, M.; LIU, B. Mining and summarizing customer reviews. In: KDD, 2004, New York-USA. Proceedings… New York-USA: ACM Press, 2004, p. 168.

KWAK, H.; LEE, C.; PARK, H.; MOON, S. What is Twitter, a social network or a news media? In: INTERNATIONAL CONFERENCE ON WWW, 19., 2010, New York. Proceedings… New York: ACM Press, 2010. p. 591.

LI, Y.M.; LI, T.Y. Deriving Marketing Intelligence over Microblogs. In: HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES, 44., 2011, Hawaii. Proceedings… Hawaii: IEEE, Jan. 2011. p. 1–10.

LIU, B. Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, n. 5, v. 1, p.1–167, May 2012.

MEIER, F.; ELSWEILER, D.; WILSON, M. More than Liking and Bookmarking? Towards Understanding Twitter Favouriting Behaviour. In: International AAAI Conference on Weblogs and Social Media, 8., 2014. Proceedings…2014.

NASCIMENTO, P.; AGUAS, R.; LIMA, D.D.; KONG, X.; OSIEK, B. Análise de sentimento de tweets com foco em notícias. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING, 1., 2012. Anais… 2012.

NICOLELLA, G.; MARQUES, J. F.; SKORUPA, L. A.; Sistema de Gestão Ambiental: aspectos teóricos e análises de um conjunto de empresas da região de Campinas, SP. São Paulo: Embrapa, 2004.

O'HARE, N.; DAVY, M.; BERMINGHAM, A.; FERGUSON, P.; SHERIDAN, P.; GURRIN, C.; SMEATON, A. F. Topic-dependent sentiment analysis of financial blogs. In: TSA, 2009, New York. Proceedings… New York: ACM Press, 2009. p. 9.

PAK, A.; PAROUBEK, P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In: CONFERENCE ON INTERNATIONAL LANGUAGE RESOURCES AND EVALUATION LREC, 70., 2010. Proceedings… ELRA, 2010. p. 1320–1326.

PANG, B.; LEE, L. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, n. 2, v. 2, 2008.

PFITZNER, R.; GARAS, A.; SCHWEITZER, F. Emotional Divergence Influences Information Spreading in Twitter. In: INTERNATIONAL AAAI CONFERENCE ON WEBLOGS AND SOCIAL MEDIA, 6., 2012. Proceedings… Ireland: Trinity College, 2012. p. 543-546.

READ, J. Using Emoticons to reduce Dependency. in Machine Learning Techniques for Sentiment Classification. In: ACL STUDENT RESEARCH WORKSHOP, 2005, Stroudsburg, PA, USA. Proceedings… Stroudsburg, PA, USA: ACM, 2005. p. 43-48.

SARMENTO, L.; CARVALHO, P.; SILVA, M.J.; DE OLIVEIRA, E. Automatic creation of a reference corpus for political opinion mining in user-generated content. In: INTERNATIONAL CIKM WORKSHOP ON TOPIC-SENTIMENT ANALYSIS FOR MASS OPINION – TSA, 1., 2009, New York. Proceedings… New York: ACM Press, 2009. p. 29.

SHARMA, A.; DEY, S. A comparative study of feature selection and machine learning techniques for sentiment analysis. In: RACS, 2012, New York. Proceedings… New York: ACM Press, 2012. p. 1-7.

SUH, B.; HONG, L.; PIROLLI, P.; CHI, E. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In: IEEE INTERNATIONAL CONFERENCE ON SOCIAL COMPUTING / IEEE INTERNATIONAL CONFERENCE ON PRIVACY, Security, Risk and Trust. 2010.

STIEGLITZ, S.; DANG-XUAN, L. Political Communication and Influence through Microblogging - An Empirical Analysis of Sentiment in Twitter Messages and Retweet Behavior. 45th In: HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES, 45., 2012, Hawaii. Proceedings… Hawaii: IEEE, Jan. 2012.

TUMITAN, D.; BECKER, K. Tracking Sentiment Evolution on User-Generated Content: A Case Study on the Brazilian Political Scene. SBBD, p. 1-6. 2013.