Análise da viralidade em eventos acadêmicos através das redes sociais

Camila dos Santos , Daniela Barreiro Claro

[1]FORMAS - Grupo de Pesquisa em Formalismos e Aplicações Semânticas
Laboratório de Sistemas Distribuídos (LaSiD) / Departamento de Ciência da Computação
Instituto de Matemática (IM)/ Universidade Federal da Bahia (UFBA)
Av. Adhemar de Barros, s/n, Ondina – Salvador – BA – Brasil
mylla.happy@gmail.com, dclaro@ufba.br

Abstract. *The present work has the main objective to classify academic events such as symposia, conferences, seminars, workshops, lectures, courses in order to analyze their spread in education concerns. Our experiments was conducted within a set of data mining algorithms and we extracted that the SMO algorithm had the best performance. As our results, we could conclude that academic seminars are the most educational event propagated over a social network.*

Resumo. *O presente trabalho tem como principal objetivo classificar os eventos acadêmicos, tais como simpósio, conferencias, seminários, workshops, palestras e cursos com o objetivo de analisar a viralidade no domínio da educação. Os experimentos foram realizados com um conjunto de três algoritmos no qual o SMO foi um dos que obtiveram melhor desempenho. Como resultados, pode-se observar que os seminários são os eventos mais propagados na rede social analisada.*

1. Introdução

Uma rede social é uma estrutura social composta por pessoas ou organizações, conectadas por um ou vários tipos de relações, que compartilham valores e objetivos comuns [Zafarani et al. 2014]. As redes sociais têm exercido um papel importante na propagação de informações, que antes eram espalhadas através do 'boca-a-boca' das pessoas. A divulgação de uma informação através de uma rede social pode, muitas vezes, ser mais efetiva do que a disseminação por antigos meios de comunicação, tais como SMS (*Short Message Sender*), emails, etc. Neste sentido, as redes sociais atualmente têm uma importante contribuição na divulgação de materiais, incluindo a disseminação de eventos.

Nos últimos anos, as redes sociais, tais como *Facebook*[1], *Twitter*[2], *Youtube*[3] têm se popularizado enormemente. Através destas redes, o compartilhamento de informações pode ocorrer de uma maneira rápida e fácil, sendo possível opinar sobre diversos assuntos, como notícias, eleições, esportes, produtos, eventos, dentre outros. O *Twitter* é um *microblogging* onde os usuários postam mensagens com até 140 caracteres, chamadas *tweet*. Os usuários podem marcar *tweets* como favorito (*favorite*), compartilhar os *tweets* (*retweet*) e responder a um *tweet* (*reply*). O *Twitter* tem sido utilizado não só por empresas, como

[1]http://facebook.com.br/
[2]https://twitter.com/
[3]http://www.youtube.com/

uma forma de marketing, mas também por jornalistas para ajudar em investigações, para encontrar pessoas que testemunharam algum incidente ou evento [Myers 2014].

Em se tratando de eventos, há uma grande disseminação de *tweets* sobre eventos, tais como: *shows*, casamentos, jogos de futebol, entre outros. Porém, observa-se que há poucos *tweets* específicos sobre eventos acadêmicos. Neste sentido, o primeiro objetivo deste trabalho foi (i) analisar a frequência dos eventos acadêmicos e consequentemente a sua viralidade, ou seja, o quanto estes eventos são disseminados. Em se tratando de viralidade, há duas possibilidades: viralidade positiva e negativa. Então, o segundo o objetivo deste trabalho foi (ii) avaliar a opinião compartilhada pelos participantes dos eventos que estavam sendo disseminados.

Assim, este trabalho propõe compreender a divulgação das informações referente aos eventos acadêmicos com o intuito de auxiliar as Universidades e seus organizadores no direcionamento de um melhor investimento para que se obtenha consequentemente uma maior divulgação do evento em questão. De certa forma, acredita-se que os resultados deste trabalho possam contribuir para uma melhoria da divulgação do evento e no crescimento da qualidade dos mesmos, já que os participantes podem dar um *feedback* mais rapidamente.

O presente trabalho está organizado como segue: na seção 2 é descrita a proposta deste trabalho. Na seção 3 os experimentos são detalhados. A seção 4 apresenta os resultados obtidos. A seção 5 discute os trabalhos relacionados e, finalmente, na seção 6 estão descritas as conclusões e os trabalhos futuros.

2. Análise dos eventos acadêmicos

O processo para analisar os eventos acadêmicos e a viralidade dos mesmos foi particionado em três principais etapas, conforme a Figura 1.

O *Twitter* foi utilizado neste trabalho em função do seu tamanho, por se tratar de um microblogging e também dada a sua facilidade de compartilhar as informações por ele processadas.

A etapa 1 corresponde à coleta dos dados, no qual foram processados e classificados como eventos acadêmicos e não-acadêmicos. Nesta classificação, a fim de obter a melhor precisão e cobertura, foram analisados três algoritmos da Mineração de Dados: *SMO* [Horst and Romeijn 2002], *Naives Bayes Multinomial* [Zafarani et al. 2014] e o *Random Tree* [Breiman 2001]. Estes algoritmos são geralmente utilizados tanto na mineração de texto como na mineração de opinião. O SMO foi selecionado por ser uma otimização do SVM [BODÓ 2009]. O *Naives Bayes Multinomial*, por considerar as frequências dos termos dentro de cada documento, ou seja, em cada *tweet*. E também foi escolhido um algoritmo de árvore de decisão, o *Random Tree*.

A etapa 2 foi responsável pela mineração de opinião dos *tweets* coletados. Os *tweets* referentes aos eventos acadêmicos da etapa anterior foram categorizados em positivo, negativo e neutro. Esta classificação também avaliou três algoritmos da Mineração de Dados com o intuito de melhor opinar em relação aos eventos acadêmicos.

A etapa 3 foi responsável por avaliar a viralidade dos *tweets* selecionados, verificando os tipos de rumores distintos, tais como *rumor positivo* e *rumor negativo*. Além

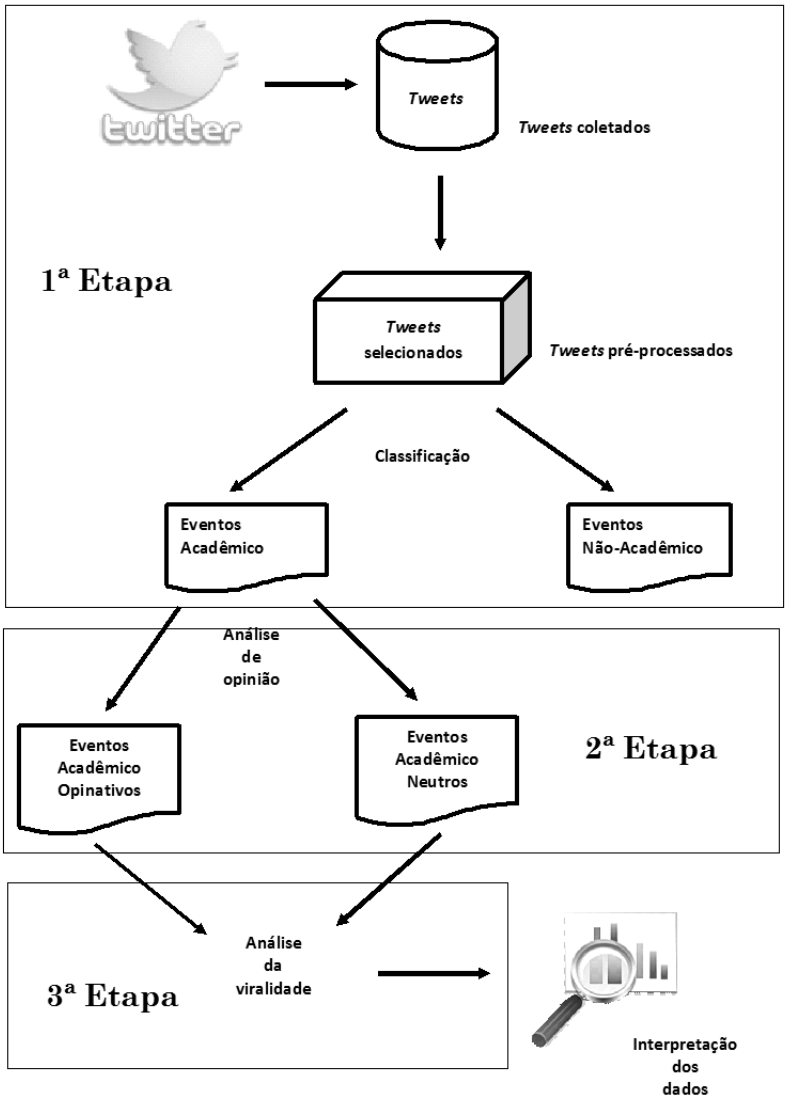

Figura 1. Etapas da Análise dos eventos.

disso, foi possível analisar a propagação do evento mencionado, culminando com a viralidade do evento acadêmico em questão.

Considerando estas três etapas, os experimentos foram conduzidos a fim de que se pudesse analisar os eventos acadêmicos.

3. Tratamento dos Dados

Nesta seção são descritos os tratamentos dos dados que foram realizados com o intuito de executar os experimentos.

3.1. Base de Dados

Os *tweets* coletados foram extraídos utilizando a API do *Twitter REST* versão 1.1, com o auxílio da biblioteca *Twython*[4] em Python, com a requisição GET *search/tweets*.

Os *tweets* foram selecionados através de uma das seguintes palavras-chaves: 'simposio', 'curso', 'seminario', 'congresso', 'palestra' e 'workshop'. Esta consulta foi realizada no período de 10/09/2014 a 16/09/2014 e somente os *tweets* em português foram coletados. As informações coletadas dos *tweets* foram: o texto e a quantidade de retweets.

A base de dados coletada foi composta por 3594 *tweets*. Os documentos que foram extraídos a cada dia foram armazenados em arquivos separados por categorias e organizados em pastas. Os *tweets* que não possuiam URLs e os *tweets* repetidos foram descartados. As URLs permitiram identificar a classe de cada evento, na etapa posterior, por isso a mesma foi utilizada como um critério de seleção para compor a base de dados.

Assim, a base de dados final continha 866 *tweets* e estes foram utilizados nos experimentos.

Devido a esta filtragem, a distribuição dos documentos nas seis categorias não ficou uniforme. A Tabela 1 apresenta a quantidade de *tweets* para cada tipo de evento que foram selecionados para utilizar no experimento.

Tabela 1. Distribuição dos *tweets* selecionados por tipo de evento.

Categoria	Quantidade
Congresso	166
Curso	62
Seminário	231
Simpósio	147
Palestra	203
Workshop	57
Total	866

3.2. Etiquetação

Com o intuito de gerar a base de treinamento para a validação dos experimentos, os *tweets* foram etiquetados manualmente. Cada um foi etiquetado como evento acadêmico ou evento não-acadêmico. A Tabela 2 apresenta a quantidade de *tweets* em cada classe.

[4]https://twython.readthedocs.org/en/latest/index.html#

Tabela 2. Quantidade de tweets etiquetados manualmente na Classificação.

Classe	Quantidade de *tweets*
Acadêmico	318
Não-Acadêmico	548
Total	866

Os *tweets* foram etiquetados como acadêmico por se tratar de um evento voltado para os universitários, confirmados pelo acesso as URLs. Os demais *tweets* que não acessavam eventos universitários foram etiquetados como evento não-acadêmico. De acordo com a Tabela 2, 37% dos *tweets* foram etiquetados como acadêmicos e 63% foram etiquetados como eventos não-acadêmicos.

3.3. Pré-processamento dos *tweets*.

Dentre as principais etapas do pré-processamento da MD [Feldman and Sanger 2006] (Mineração de Dados), destacam-se: Remoção das StopWords, Tokenização e Limpeza dos dados. Todas estas etapas foram realizadas na base de dados coletada.

Stopwords adaptada. Com o intuito de processar cada *tweet*, os mesmos foram pré-processados a fim de destacar as palavras mais relevantes para a classificação. Uma lista de *stopword* para a Língua Portuguesa foi adaptada de [do SnowBall 2002] e novas palavras, específicas do Twitter, foram adicionadas. Assim, obteve-se um novo conjunto de 309 palavras,conforme Tabela 3 .

Tabela 3. Lista de *stopwords* adaptada ao Twitter.

Palavras acrescentadas
hoje, amanha, ontem, que, janeiro, fevereiro, marco, abril, maio, etc, junho, julho, agosto, setembro, outubro, novembro, dezembro, segunda-feira, terca-feira, quarta-feira, quinta-feira, sexta-feira, segunda, terca, quarta, quinta, sexta, sabado, domingo, youtube, yahoo, rt, retweets, http

Tokenização e Limpeza dos dados A tokenização foi realizada com o auxílio da ferramenta WordTokenizer [of Waikato 2015], onde os termos foram tokenizados. Em seguida, houve a limpeza dos dados. A limpeza ocorreu através da remoção de números e de termos referentes aos usuários do Twitter, que são palavras que começam com o símbolo @, como por exemplo @ufg_oficial, usuário da Universidade Federal de Goiás e @Simpotur, usuário da organização do Simpósio de Turismo.

Uma vez concluído o tratamento dos dados, os experimentos foram executados na base de treinamento e na base de teste.

4. Experimentos

Os experimentos foram divididos em três etapas, conforme descrito na seção 2. Assim foi possível avaliar cada uma das etapas propostas neste trabalho.

4.1. 1a etapa dos experimentos

A partir do corpus coletado, o conjunto de treinamento e o conjunto de teste foram separados. Dos 866 *tweets*, 606 (70%) documentos foram selecionados para o conjunto de treinamento e 260 (30%) para o conjunto de teste.

Os algoritmos classificadores *SMO* [Horst and Romeijn 2002], *Naives Bayes Multinomial* [Zafarani et al. 2014] e o *Random Tree* [Breiman 2001] foram utilizados para testar o desempenho e analisar os melhores resultados. Em seguida, os algoritmos foram avaliados sob o conjunto de teste.

4.2. 2a etapa dos experimentos

Uma vez que os eventos foram classificados em acadêmicos, a 2a etapa corresponde à análise da opinião destes eventos. Os *tweets* opinativos foram classificados em *positivo* ou *negativo*, e os sem opinião foram classificados como *neutro*. Os *tweets* foram classificados como positivo, por expressarem boas opiniões referente ao evento, negativo os que expressaram opiniões ruins e neutro os que não expressaram nenhum tipo de opinião.

Nesta segunda etapa, também foi gerado um conjunto de treinamento e um conjunto de testes. Para o conjunto de treinamento foram etiquetados manualmente os *tweets* acadêmicos em: positivo, negativo ou neutro, como mostra a Tabela 4.

Tabela 4. Quantidade de tweets etiquetados manualmente na Mineração de Opinião.

Classe	Quantidade de *tweets*
Positivo	37
Negativo	6
Neutro	270

Os *tweets* foram classificados manualmente como positivo por expressarem uma opinião boa em relação ao evento. A opinião boa em relação ao evento se referiu à presença de palavras, tais como: *bom, feliz, alegre, legal*. Os outros *tweets* foram negativos por apresentarem uma opinião ruim sobre o evento, no qual continham as palavras: *ruim, nervoso, triste, chato*. Os eventos foram classificados como neutros por não expressarem nenhum tipo de opinião.

A Tabela 5 mostra exemplos de *tweets* de eventos acadêmicos para cada sentimento. Foi observado que *tweets* neutros apresentam apenas informações sobre o evento acadêmico.

Tabela 5. Exemplo de *tweets* de eventos acadêmicos para cada sentimento.

Sentimento	Tweet
Positivo	com o Dr. Augusto Cury aqui em Mococa, bom demais, palestra sensacional
Negativo	#Brasilia #seminario #PIBIC #nervosodemais @ICMBio
Neutro	A IBEFGV promove hoje a palestra Lideranca Essencial para abertura das aulas do curso de POS ADM

Os mesmos três algoritmos: *SMO* [Horst and Romeijn 2002], *Naives Bayes Multinomial* [Zafarani et al. 2014] e o *Random Tree* [Breiman 2001] foram utilizados para o conjunto de treinamento. Em seguida, os mesmos foram validados com o conjunto de teste.

4.3. 3a etapa dos experimentos

Nessa terceira etapa do trabalho, os *tweets* classificados quanto à opinião e à quantidade de *retweets* foram analisados. O principal intuito era avaliar a viralidade destes eventos, considerando a opinião emitida em cada *tweet*. Assim, a quantidade de *tweets* classificados por classe (positivo e negativo) foi somada, incluindo a quantidade de *retweets* para cada tipo de evento acadêmico.

4.4. Medidas de Desempenho

As medidas de desempenho servem para melhor subsidiar uma comparação entre métodos utilizados[Santos 2013]. A matriz de confusão é o resultado dos classificadores após a execução nos conjuntos de treinamento e de teste, como é apresentada na Tabela 6. Quanto maior os valores na diagonal principal da matriz (VP e VN) e menor os valores na diagonal secundária (FN e FP) melhor é o método avaliado [Oliveira 2012].

Tabela 6. Matriz de confusão 2 x 2.

		Valor Verdadeiro (confirmado por análise)	
		Positivos	Negativos
Valor Previsto (predito por teste)	Positivos	VP (Verdadeiro Positivo)	FP (Falso Positivo)
	Negativos	FN (Falso Negativo)	VN (Verdadeiro Negativo)

Com o intuito de medir a qualidade da predição, utiliza-se a acurácia, a precisão, cobertura e *Medida-F*[Feldman and Sanger 2006].

A acurácia (Equação 1) ou taxa de acerto de uma categoria é a porcentagem dos documentos classificados corretamente. A precisão (Equação 2) é a porcentagem dos documentos classificados corretamente entre todos os documentos classificados[Feldman and Sanger 2006]. A cobertura (Equação 3) é a porcentagem dos documentos classificados corretamente entre todos os documentos corretos e a *Medida-F* (Equação 4) é uma medida harmônica entre precisão e a cobertura.

Assumindo valores no intervalo [0,1], o valor 0 corresponde a nenhum documento relevante recuperado, enquanto que o valor 1 corresponde a todos os documentos recuperados são relevantes e todos documentos relevantes foram recuperados [Zhang and Zhang 2009].

$$acuracia = \frac{VP + VN}{VP + VN + FP + FN} \tag{1}$$

$$precisao = \frac{VP}{VP + FP} \tag{2}$$

$$cobertura = \frac{VP}{VP + FN} \tag{3}$$

$$\textit{F-measure} = \frac{2 * precisao * cobertura}{precisao + cobertura} \tag{4}$$

Outra alternativa é o uso dos gráficos ROC, que são gráficos bidimensionais, onde o eixo Y tem o valor referente a quantidade de verdadeiros positivos e o eixo X tem o valor dos falsos positivos [Silva 2006].

Uma forma de comparar curvas ROC é utilizando o cálculo da área sob a curva (AUC), que é uma medida escalar variável no intervalo entre 0 e 1. Os resultados com maiores valores de AUC são considerados os melhores [Gomes 2006].

5. Resultados

Os resultados serão apresentados por etapas, de acordo com a seção 2 deste trabalho, em relação a cada experimento realizado.

5.1. 1a etapa dos Resultados

Observou-se que o desempenho dos três algoritmos utilizando o conjunto de treinamento foi similar, visto que as medidas obtidas foram bem semelhantes, conforme Tabela 7.

Tabela 7. Resultados da classificação do conjunto de treinamento.

Algoritmos	Acurácia	Precisão	Cobertura	F-Measure	Área ROC
SMO	0.998	0.998	0.998	0.998	0.998
Naives Bayes Multinomial	0.993	0.993	0.993	0.993	1
Random Tree	0.998	0.998	0.998	0.998	1

Em seguida, o mesmo experimento foi realizado com o conjunto de teste, conforme Tabela 8. De acordo com os resultados obtidos, constatou-se que o algoritmo SMO obteve melhores medidas de desempenho.

Tabela 8. Resultados da classificação do conjunto de teste.

Algoritmos	Acurácia	Precisão	Cobertura	F-Measure	área ROC
SMO	**0.692**	**0.681**	**0.692**	**0.677**	0.637
Naives Bayes Multinomial	0.611	0.676	0.612	0.616	**0.693**
Random Tree	0.665	0.651	0.665	0.651	0.612

Assim, na classificação entre eventos acadêmicos e não-acadêmico, referente a primeira etapa deste trabalho, o algoritmo SMO obteve melhores medidas (acurácia, precisão, cobertura e F-measure), embora o Naives Bayes Multinomial tenha obtido a maior área ROC.

5.2. 2a etapa dos Resultados

Na segunda etapa dos resultados, o objetivo foi avaliar a classificação da opinião dos *tweets*. Considerando o conjunto de treinamento e os algoritmos utilizados apresentados na Tabela 9, observou-se que os dois algoritmos, SMO e o *Random Tree* obtiveram desempenhos semelhantes na fase de construção do modelo.

Tabela 9. Resultados do conjunto de treinamento da Mineração de Opinião.

Algoritmos	Acurácia	Precisão	Cobertura	F-Measure	área ROC
SMO	1	1	1	1	1
Naives Bayes Multinomial	0.991	0.992	0.991	0.991	1
Random Tree	1	1	1	1	1

Considerando o conjunto de testes, conforme a Tabela 10, o SMO obteve os melhores resultados, similar a etapa 1.

O classificador de Naive Bayes Multinomial apresentou uma taxa de acerto de 27.6 % na classificação de 94 *tweets* presentes no conjunto de teste. Destes 94 documentos, 9 são *tweets* positivos, 2 *tweets* negativos e 83 *tweets* neutros. A Tabela 11 exibe os resultados da classificação para cada categoria de *tweets* pertencentes ao conjunto de teste.

O algoritmo *Random Tree* classificou corretamente 81.9% dos 94 *tweets* presentes no conjunto de teste, conforme mostra a Tabela 12.

Já o algoritmo SMO teve uma taxa de acerto de 88.3 % dos 94 *tweets* presentes no conjunto de teste, conforme mostra a Tabela 13.

Tabela 10. Resultados do conjunto de teste da Mineração de Opinião.

Algoritmos	Acurácia	Precisão	Cobertura	F-Measure	área ROC
SMO	**0.883**	0.835	**0.883**	**0.845**	0.54
Naives Bayes Multinomial	0.276	**0.863**	0.277	0.39	**0.671**
Random Tree	0.819	0.792	0.819	0.805	0.504

Tabela 11. Total de *tweets* classificados como positivo, negativo e neutro pelo classificador Naive Bayes Multinomial.

Categoria	Quantidade	Positivo	Negativo	Neutro
Positivo	9	3	6	0
Negativo	2	0	1	1
Neutro	83	13	48	22

Na mineração de opinião foi verificado que o algoritmo SMO obteve melhor desempenho no conjunto de teste em comparação com outros algoritmos. Foi observado também que a maioria dos usuários que postaram as mensagens eram referentes a perfis de universidades e dos organizadores dos eventos acadêmicos.

5.3. 3a etapa dos Resultados

A terceira etapa dos resultados corresponde a análise da viralidade dos eventos acadêmicos, conforme Tabela 14. As seguintes características da viralidade foram utilizadas: rumor positivo, rumor negativo e propagação. Tanto o rumor positivo quanto o negativo foram analisados baseados na mineração de opinião, enquanto que a propagação foi de acordo com a quantidade de *retweets* de cada *tweet*.

Os tweets classificados como positivo foram mapeados como rumor positivo e os da classe rumor negativo foram mapeados para rumor negativo. Em seguida, foi calculada a soma da quantidade de *retweets* dos eventos acadêmicos. Eventos do tipo congresso tiveram um maior rumor positivo enquanto os eventos do tipo seminário foram detectados alguns rumores negativos. Porém, embora os rumores tenham sido negativos para os seminários, houve uma maior propagação em relação aos demais.

Com base nos experimentos realizados em relação às três etapas, pôde-se observar que o algoritmo SMO mostrou-se mais eficiente. Vale ressaltar também que o algoritmo apresenta valores maiores para acurácia, precisão, cobertura e Medida-F, porém apresenta valores menores para área ROC. Isso ocorre, devido o algoritmo SMO ter um valor menor referente a quantidade de verdadeiros positivos do que o valor dos falsos positivos.

5.4. Trabalhos Relacionados

Com o intuito de melhor posicionar a relevância do presente trabalho, alguns autores tem abordado uma análise em redes sociais, especificamente utilizando o Twitter.

Em [Avvenuti et al. 2014], os autores propõem uma arquitetura para implementar um cenário real no qual eles utilizam o *Twitter* para detectar terremotos no território

Tabela 12. Total de *tweets* classificados como positivo, negativo e neutro pelo classificador Random Tree.

Categoria	Quantidade	Positivo	Negativo	Neutro
Positivo	9	1	0	8
Negativo	2	0	0	2
Neutro	83	7	0	76

Tabela 13. Total de *tweets* classificados como positivo, negativo e neutro pelo classificador SMO.

Categoria	Quantidade	Positivo	Negativo	Neutro
Positivo	9	1	0	8
Negativo	2	0	0	2
Neutro	83	1	0	82

Tabela 14. Viralidade dos *tweets* por categoria.

Característica	Congresso	Curso	Palestra	Seminário	Simpósio	Workshop	Total
Rumor Positivo	**12**	2	10	8	4	0	36
Rumor Negativo	1	0	2	**3**	1	0	7
Propagação	14	6	21	**60**	48	1	150

italiano. Utilizaram o algoritmo J48 [Quinlan 1993] com 10-fold cross validation. Os resultados mostraram a alta habilidade do sistema de detectar terremotos com magnitude maior ou igual a 3.5 Richter com apenas 10% de falsos positivos.

O *TweetMining* trata-se de um sistema no padrão MVC, utilizando a arquitetura cliente-servidor, que coleta, pré-processa, treina e classifica os *tweets* referentes a cidade de Campinas e apresenta os dados através de gráficos. Utilizando a implementação do *SVMLearn* [BODÓ 2009] para a mineração de opinião dos dados [Sousa 2012].

O Twitter também tem sido utilizado no âmbito educacional. Em [Falcao et al. 2011] os autores propõem o uso do *Twitter* para a divulgação de informações de ações, projetos, eventos, cursos, editais e notícias da Universidade de Pernambuco, a fim de aproximar a universidade da sociedade. Ressaltam a importância de estimular a participação dos alunos e professores para que criem conteúdos exclusivos para serem compartilhados.

Em [Benevetuno 2013] o objetivo é identificar as áreas de atuação de pesquisadores combinando técnicas de mineração de textos com a análise de redes sociais. Utilizando currículos da Plataforma Lattes, o autor extraiu características baseada em Mineração de Texto, utilizando a frequência relativa das palavras dos títulos dos artigos publicados em periódicos e baseada na Análise de Redes Sociais, pela porcentagem dos vizinhos pertencentes a cada grande área. Para os experimentos, foi utilizado o *Rotation Forest* [Rodriguez 2006]. Apesar das técnicas empregadas serem relativamente simples, os resultados obtidos foram bastante satisfatórios, atingindo taxas de acerto superiores a 90% para a identificação das grandes áreas (dentre as 8 disponíveis); superiores a 84% para a identificação de áreas (dentre 76); e de 59,77% para a identificação de sub-áreas (dentre 443). Verificou-se que a combinação simples da técnica de mineração de textos utilizada com a análise da vizinhança de nível dois obteve melhores resultados do que aqueles produzidos pelas técnicas usadas individualmente. Para o uso de mineração de texto, observou-se que usar dados com mais de 9 anos para a identificação das grandes áreas resultou, em muitos casos, na diminuição da taxa de acerto.

Em [Silva et al. 2010], o autor apresenta uma metodologia para a análise de propagação de URLs no *Twitter* que foi implementada no portal Observatório da Web, com o objetivo de analisar a repercussão de grandes eventos na web. Entre as técnicas utilizadas estão a identificação de palavras-chave e tópicos associados as URLs, a análise de propagação de outros itens de informação e a identificação de usuários com grande capacidade de propagar informação no *Twitter*.

Já em [Guerini et al. 2011], os autores apresentam as características da viralidade (apreciação, propagação, rumor simples, rumor positivo, rumor negativo, aumentar a discussão e controversalidade) dos conteúdos nas redes sociais online, com a análise no *Digg*[10]. Mostrando que o fenômeno da viralidade pode ser predita considerando apenas o alcance do conteúdo e a interdependência entre suas características.

Os trabalhos acima não analisaram juntamente a opinião e a viralidade das redes sociais para eventos acadêmicos. Em [Falcao et al. 2011] foi feita uma análise geral do uso do *Twitter* em universidades, porém não houve uma análise quantitativa da propagação de informações acadêmicas no *Twitter*. Neste trabalho, é proposto uma análise quantitativa da viralidade de eventos acadêmicos do *Twitter*, incluindo os respectivos rumores.

6. Conclusão e Trabalhos Futuros

Neste trabalho foi proposto um conjunto de etapas para a análise da viralidade de eventos acadêmicos postados no *Twitter*. Essas etapas foram divididas em três: a classificação dos *tweets* em eventos acadêmicos e não-acadêmicos, a mineração de opinião dos eventos acadêmicos e a análise da viralidade dos mesmos. Com o propósito de validar essa proposta foram realizados experimentos com os dados coletados.

Comparando-se os resultados obtidos, pode-se concluir que o algoritmo SMO possui uma eficiência satisfatória com relação às medidas de acurácia, precisão, cobertura e Medida-F, tanto para a classificação quanto para a mineração de opinião. Quanto a viralidade, os seminários são mais propagados no meio acadêmico.

Diversas dificuldades foram encontradas no decorrer do trabalho, como por exemplo, o tamanho reduzido dos *tweets*, os ruídos nos dados e a análise da opinião dos eventos. Os possíveis trabalhos futuros são:

- Utilização de outras técnicas como o *n-gram* na classificação, para uma possível comparação entre os métodos aqui utilizados.
- Realização da etiquetação automatizada dos *tweets*.
- Inclusão de outras modalidades de eventos acadêmicos, tais como: mesa-redonda, painel, fórum, conferência, entre outros.
- Utilização de outros algoritmos na fase da classificação, tais como: *Maximum Entropy*, J48 [Quinlan 1993] e *Rotation Forest*.
- Utilização das outras características da viralidade, tais como: apreciação, discussão e controversalidade.

Referências

Avvenuti, M., Cresci, S., Marchetti, A., Polla, M., and Tesconi, M. (2014). In *Earthquake emergency management by Social Sensing*. IEEE. Último acesso em dezembro de 2014.

Benevetuno, F. (2013). Combinando mineração de textos e análise de redes sociais para a identificação das Áreas de atuação de pesquisadores. Último acesso em dezembro de 2014.

[10]http://digg.com/

BODÓ, Z. (2009). Support vector machine library in java. Último acesso em dezembro de 2014.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

do SnowBall, D. (2002). Portuguese stop word list. Último acesso em dezembro de 2014.

Falcao, L., Junior, C., and Arruda, F. (2011). Proposta de uso do twitter como ferramenta de informação e conhecimento na universidade de pernambuco. *Anais do XXII SBIE - XVII WIE*.

Feldman, R. and Sanger, J. (2006). *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, New York, NY, USA.

Gomes, D. d. A. d. O. (2006). Detecção de instrusão em redes de computadores utilizando classificadores one-class. Último acesso em dezembro de 2014.

Guerini, M., Strapparava, C., and Ozbal, G. (2011). Exploring text virality in social networks. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 506–509.

Horst, R. and Romeijn, H. E. (2002). *Handbook of global optimization*, volume 2. Springer.

Myers, P. (2014). Como usar twitter para encontrar pessoas em notícias urgentes. Último acesso em dezembro de 2014.

of Waikato, U. (2015). Weka 3 – machine learning software in java. Último acesso em abril de 2015.

Oliveira, S. R. d. M. (2012). Medidas para avaliação de regras e de modelos de classificação. Último acesso em dezembro de 2014.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Rodriguez, J. J. (2006). Rotation forest: A new classifier ensemble method. *IEEE TRANS. PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, pages 1619–1630.

Santos, F. (2013). Mineração de opinião em textos opinativos utilizando algoritmos de classificação. Último acesso em dezembro de 2014.

Silva, A., Mourao, F., Simoes, L., and Veloso, A. (2010). Análise de padrões de propagação no twitter.

Silva, F. C. (2006). Análise roc. Último acesso em dezembro de 2014.

Sousa, G. L. S. d. (2012). Tweetmining: Análise de opinião contida em textos extraídos do twitter. Último acesso em dezembro de 2014.

Zafarani, R., Abbasi, M., and Liu, H. (2014). *Social Media Mining: An Introduction*. Cambridge University Press.

Zhang, E. and Zhang, Y. (2009). F-measure. In LIU, L. and ÖZSU, M., editors, *Encyclopedia of Database Systems*, pages 1147–1147. Springer US.