SOMTOOLS UMA FERRAMENTA PARA ANÁLISE DE OPINIÃO E SENTIMENTO NO AMBIENTE TWITTER

Leila Weitzel
Universidade Federal do Sul e Sudeste do Pará, Faculdade de Computação - UNIFESSPA
Marabá, Brasil
lmartins@ufpa.br

Raul Freire Aguiar
Universidade Federal do Sul e Sudeste do Pará, Faculdade de Computação - UNIFESSPA
Marabá, Brasil
raulfreire.si@gmail.com

Walter Fernando García Rodriguez
Universidade Federal do Sul e Sudeste do Pará, Faculdade de Computação - UNIFESSPA
Marabá, Brasil
wfernando.grodriguez@gmail.com

Marcela Gomes Heringer
Universidade Federal do Sul e Sudeste do Pará, Faculdade de Computação - UNIFESSPA
Marabá, Brasil
marcelag.heringer@gmail.com

Resumo - Um tipo de informação relevante disponível na Web são as opiniões expressas no conteúdo que é gerado pelo próprio usuário, postadas no Twitter, Facebook e em outras redes sociais. Análise de sentimento e mineração de opinião faz parte do campo de estudo que analisa as opiniões das pessoas, sentimentos, avaliações, atitudes e etc. Recentemente, tem despertado grande interesse, tanto na academia quanto nas organizações devido ao desenvolvimento de aplicações úteis. Neste trabalho avaliou-se um conjunto de tweets da linha de tempo de usuários do Twitter. O conjunto de dados foi analisado com base em duas abordagens diferentes: a análise léxica e análise sintática. Foi realizado também um estudo comparativo de modo a verificar se existe correlação entre o sentimento das mensagens e a reputação destes usuários. Os achados experimentais com ambas as abordagens foram similares, significando que a remoção stopword não altera os resultados. Nas seções seguintes são detalhadas a aplicação e as limitações desta pesquisa

Palavras-chave- Sentimento e Mineração de Opinião, Twitter, SentiWordNet

Abstract- One of the important types of information on the Web is the opinions expressed in the user generated content, posted on Twitter, Facebook and others social networks. Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and etc. Recently, it has attracted great interest both in academia and in industry due to its useful potential applications. In this paper we evaluated a set of tweets from a timeline of Twitter users. We analyzed the dataset based on two different approaches: lexical analysis and syntactic analysis. We also perform a comparative study to find out if there is a correlation between user reputation and user sentiment. The experimental finding with both approaches shown similar results, which means, that does not make difference between stopword removal or not. Further, the paper exemplifies the applications of this research and its limitations.

Keywords- Sentiment and opinion mining, Twitter, SentiWordNet.

I. INTRODUÇÃO

Nas últimas décadas, o rápido crescimento de Blogs, Fóruns e Redes Sociais na Web tornaram visíveis "opiniões" a respeito dos mais diversos assuntos. Esses ambientes transformaram-se em verdadeiras plataformas de informação e comunicação, registrando publicamente pensamentos, opiniões e sentimentos sobre tudo [1]. Esse ambiente propício estimulou o estudo e desenvolvimento de sistemas visando avaliar opiniões automaticamente, surgindo assim a área de Análise de Sentimentos (AS) [2].

AS é o estudo de opiniões, sentimentos e emoções expressas em textos [2]. Existe uma série de estudos nessa área, como por exemplo, a extração de elementos relacionados à opinião [3], a classificação da opinião (positivo, negativo ou neutro) [4], comparação de sentenças quanto a suas opiniões [5] entre outros sobre o mesmo tema. O comportamento humano está condicionado, na maioria das vezes, em como os outros veem e avaliam o mundo. A opinião é um conceito relacionado aos sentimentos, atitudes e emoções e é dentro deste escopo que se insere esta pesquisa, cujo principal objetivo é analisar a polaridade dos sentimentos das mensagens que são postadas no ambiente Twitter. O Twitter é uma Rede Social e um *Microblogging* permitindo que seus usuários troquem mensagens rápidas. Os usuários do Twitter selecionados para o estudo de caso é um conjunto de usuários estudados em [6].

Nesse estudo [6], avaliaram a reputação de fontes de informação no domínio da saúde utilizando como base de dados o Twitter. Em especial os autores utilizaram das premissas da Análise de Redes Sociais (ARS). Como resultado, foi apresentado um *rank* (lista ordenada) dos 1232 usuários em função da reputação que estes apresentaram. Desta forma pretende-se avaliar a polaridade das mensagens que são propagadas por estes usuários. A abordagem adotada nesta análise visa testar a seguinte hipótese: "existe correlação entre a reputação e a polaridade dos sentimentos expressos por estes usuários".

O trabalho está estruturado da seguinte forma: a seção 2 são apresentados os trabalhos correlatos; a seção 3 apresenta o ambiente da pesquisa e a fundamentação teórica, na seção 4 a metodologia adotada; a seção 5 apresenta os resultados iniciais da pesquisa. E por fim, as conclusões, trabalhos futuros e referências.

II. TRABALHOS CORRELATOS

Nesta seção serão apresentados os principais trabalhos correlatos que nortearam esta pesquisa.

O estudo apresentado por [7] observou que o desempenho na classificação de opinião em mensagens do Twitter está relacionado a atribuição de pesos semânticos associadas às métricas propostas em [5].

Em [8] os autores utilizam o SentiWordNet levando-se em consideração o contexto semântico do texto, e refinada pela medida NGD [9].

E por fim, em [1] foi apresentado um estudo comparativo entre os oito métodos propostos na literatura: LIWC, Happiness Index, SentiWordNet, SASA, PANAS-t, Emoticons, SenticNet e SentiStreng. A pesquisa tem como objetivo avaliar o desempenho da classificação em função das métricas: abrangência (medindo a fração de mensagens capturadas por um método) e concordância (medindo a concordância entre a polaridade entre os métodos utilizando uma base rotulada).

III. METODOLOGIA

A. *Ambiente da pesquisa*

Twitter tem como objetivo a atualização de *status* através do envio de mensagens curtas (de até 140 caracteres) chamadas de *tweet*. Esse ambiente é propício para estudo na área de AS permitindo coletar e analisar os dados em grande escala [10].

De acordo com [2], é possível categorizar informações de texto como fatos ou opiniões. Para o autor, um fato pode ser dito como uma informação de caráter objetivo sobre alguma entidade, algum evento, algum dado ou alguma de suas propriedades. Considerando-se que a opinião apresenta um sentido subjetivo expresso por algum indivíduo ou grupo. O objetivo da AS não é determinar sobre qual tópico ou tema o documento trata (como realizado em técnicas convencionais de classificação de texto), mas sim em descobrir qual a opinião expressa no documento e, classificar a sua polaridade [11].

B. *Ferramentas utilizadas e desenvolvidas*

Foi utilizado o algoritmo (mais precisamente a classe genérica) que calcula a polaridade dos textos da ferramenta SentiWordNet [12]. O algoritmo proposto por [12], vem sendo amplamente utilizado na literatura, o que permite fazer estudos comparativos entre o nosso e outros sobre o mesmo tema. Esta ferramenta é baseada em um dicionário léxico em inglês chamado WordNet [13]. O Wordnet é composto por grupos léxicos, tais como: adjetivos, substantivos, verbos e outras classes gramaticais dentro de um conjunto de sinônimos, chamados de *synsets*. O algoritmo associa três *scores* com *synsets* do dicionário WordNet para indicar o sentimento do texto: positivo, negativo e objetivo (neutro) [12]. De acordo com [1], os *scores*, são valores entre [0, 1] obtidos utilizando-se o método de aprendizagem de máquina semi-supervisionado.

Conforme dito anteriormente, o conjunto de dados é proveniente dos achados em [6]. A listagem completa contém 1232 usuários ordenados em função da sua reputação. Nesta pesquisa inicial, foram selecionados apenas os 46 primeiros da listagem como usuários semente. A nomenclatura vem do método *Snowball que* consiste em coletar um grafo de uma rede social online seguindo uma abordagem de busca em largura. A coleta inicia-se a partir de nodo semente. Ao coletar a lista de vizinhos desse nodo, novos nodos são descobertos e então coletados no segundo passo, que só termina quando todos os nodos descobertos no primeiro passo são coletados. Em nossa pesquisa utilizamos esses usuários semente para coletar os tweets que foram postados em sua linha de tempo (timeline).

Foi implementado um *web crawler*, cujo propósito é fazer uma busca sistemática para extrair os *tweets* da *timeline* dos 46 usuários. O período de coleta abrange os meses de agosto/setembro de 2013. Foram coletados 86622 *tweets*, armazenados em formato texto. Foram desenvolvidas também ferramentas para a:

(i) **Análise Sintática** (*parsing* em inglês) dos *tweets* para a retirada de caracteres especiais tais como /, %, $ e etc. São aceitas apenas as letras de a-z, os números 0-9 e os símbolos @ (arroba), # (tralha) e o apóstrofe;

(ii) **Análise Léxica** (conhecida como *tokenização*) tem como objetivo decompor o texto em unidades estruturais menores, em nosso caso as unidades estruturais são as palavras.

(iii) **Remoção das *stopword***, ***s***egundo [14], uma *stopword* pode ser traduzida "palavra vazia", elas aparecem em praticamente todos os documentos, ou na maioria deles, por isso não são capazes de colaborar na análise da polaridade do texto.

A Figura 1 ilustra a abordagem metodológica adotada, dividida em três fases: **Primeira Fase** representa a fase extração dos *tweets*; a **Segunda Fase** diz respeito ao tratamento dos *tweets* coletados, A **Terceira Fase**, os conjuntos de teste gerados são submetidos à avaliação percorrendo as classes gramaticais suportadas pelo SentiWordNet, e ao final é gerado o *score* de cada *tweet* para cada usuário.

Para avaliar a metodologia proposta foram gerados dois conjuntos de teste, nomeados de SemSW sem as *stopword* e caracteres especiais e outro ComSW com as *stopword*. Deve-se ressaltar também que foram retirados do conjunto da amostra todos os *retweets* presentes na

timeline destes usuários para que ficassem apenas as mensagens postadas por eles. O *retweet* é uma mensagem que foi recebida e depois foi encaminhada, tem o mesmo significado semântico de um *reply* em uma mensagem de correio eletrônico.

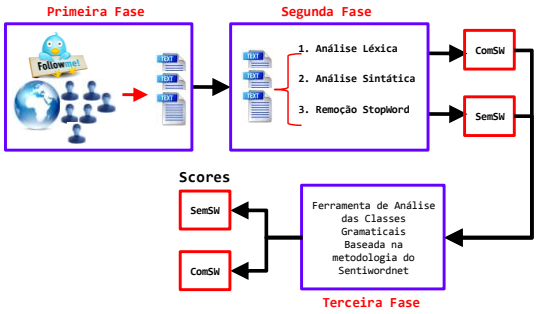

Figura 1: Visão geral da abordagem metodológica adotada

IV. RESULTADOS

Nesta seção serão apresentados os resultados da análise exploratória dos dados. Foram coletados 86622 *tweets* ao total. A média de *tweets* por usuários foi de aproximadamente 1882.

A Tabela 1 mostra a média e desvio padrão verificados para os *scores* das abordagens com e sem *stopword* de toda a amostra, ou seja, de todos os 86622 *tweets*.

TABELA 1: MÉDIA E DESVIO PADRÃO DOS 86622 *SCORES* COM E SEM *STOPWORD*

Scores ComSW	Scores SemSW
Média	**Média**
0,2669	0,3228
Desvio Padrão	**Desvio Padrão**
0,6114	0,6233

O primeiro teste aplicado foi o Teste *Kolmogorov-Smirnov*. O teste destina-se a averiguar se uma amostra pode ser considerada como proveniente de uma população com uma determinada distribuição. A primeira hipótese a ser testada para fazer inferências para uma população a partir de uma amostra, é de que esta seja aleatória. Em nosso caso para verificar se conjunto de dados (86622 *scores*) apresenta distribuição normal (Figura 2 e 3). Assim, com nível de confiança de 95% não se pode rejeitar a hipótese nula, ou seja, de que as duas distribuições são normais. A qualidade das inferências feitas por estes métodos depende de quão próxima é a população em estudo da normal. Isso quer dizer que os valores da amostra se encontram em torno da média.

A Figura 4 mostra o diagrama de pizza do percentual da polaridade verificada no conjunto de dados. A maior porcentagem diz respeito ao sentimento neutro, ou seja, sentimento considerado objetivo. Somando-se a polaridade muito positiva, positiva e fraco positivo tem-se aproximadamente 48% de positividade.

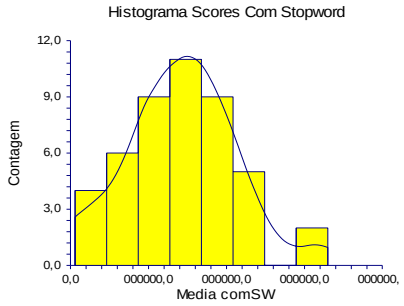

Figura 2: histograma da distribuição dos *Scores* com *stopword*

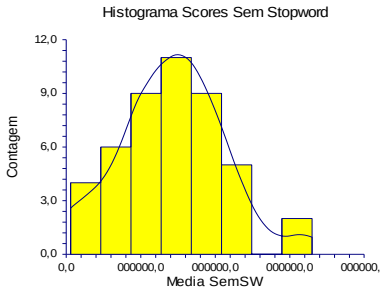

Figura 3: histograma da distribuição dos *scores* sem *stopword*

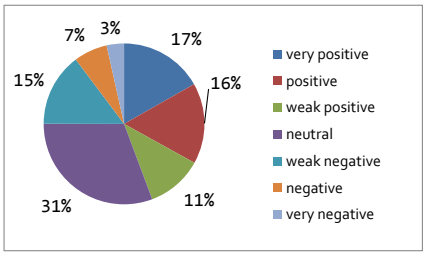

Figura 4: Percentual total da polaridade verificada na amostra

A hipótese que se quer testar com esta pesquisa é se existe correlação entre o *Rank* proposto por [6] e os *scores* (polaridade das mensagens). Para validar esta hipótese utilizou-se a análise de regressão para examinar se existe uma relação linear bivariada entre o *Rank* e os *Scores* das abordagens sem e com *stopword*. Assim sendo, foi calculada a média de todos os *scores* (positivos, negativos, fraco negativo, fraco positivo, neutro etc.) de cada usuário e esta média será avaliada com o *Rank* (reputação).

A Figura 5 e 6 mostram os resultados desta análise. Com nível de confiança de 95%, não foi possível ajustar uma relação entre estas variáveis, indicando que não existe linearidade entre elas.

Como a correlação de *Pearson* requer a suposição de que a relação entre as variáveis seja linear optou-se então pelos testes de correlação de *Spearman Rho* e de *Kendall*

Tau. Os testes de Correlação de *Spearman Rho* e de *Kendall Tau*, são teste não-paramétricos que podem ser aplicadas a listas ordenadas e que ao contrário do Teste de *Pearson* não requer linearidade entre as variáveis.

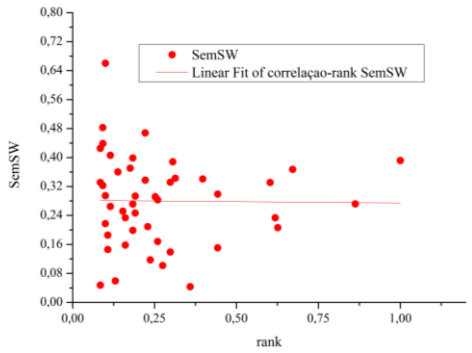

Figura 5: Diagrama de dispersão da regressão linear entre as variáveis *Rank* e *Scores* sem *stopword*

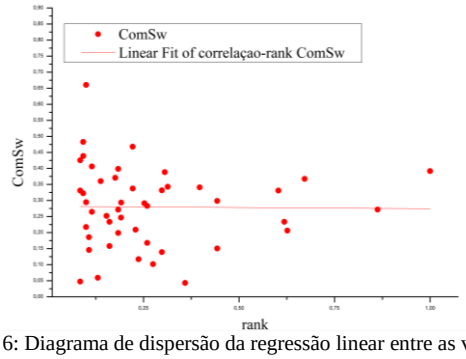

Figura 6: Diagrama de dispersão da regressão linear entre as variáveis *Rank* e *Scores* com *stopword*

TABELA 2: RESULTADO DAS CORRELAÇÕES NÃO-PARAMETRICAS

		Rank	**Média ComSW**	**Média SemSW**
Kendall's tau_b	**Rank**	100%	-7%	-7%
	Média ComSW	-7%	100%	100%
	Média SemSW	-7%	**100%**	100%
Spearman's Rho	**Rank**	100%	-10%	-10%
	Média ComSW	-10%	100%	100%
	Média SemSW	-10%	**100%**	100%

Os resultados dos testes de correlação *Spearman Rho e Kendall Tau* estão ilustrados na Tabela 2. Tanto o teste unicaudal quanto o teste bicaudal apresentaram os mesmos resultados e por este motivo é apresentado apenas um deles. É interessante perceber que as médias (SemSw e ComSW) estão fortemente correlacionadas entre si, correlação de 100%. O mesmo não pode ser observado entre o *Rank* e as médias dos *scores*, sugerindo que não existe uma relação linear entre o *Rank* proposto por [6] e o grau de polaridade da mensagem que é postada por um usuário.

Calculou-se a correlação de *Pearson* entre os 86622 *scores* com e sem *stopword*. Verificou-se uma correlação de 86%, concordando com o resultado encontrado (forte correlação) entre as médias dos *scores*.

V. CONCLUSÃO

Nesta pesquisa buscou-se analisar a polaridade dos sentimentos das mensagens que são postadas na rede social Twitter. A abordagem adotada visou testar a hipótese de que existe uma correlação entre a reputação de determinado usuário e a polaridade de suas opiniões, ou ainda, que o sentimento expresso em suas mensagens poderia influenciar a sua reputação.

Os resultados evidenciaram que existe uma forte correlação entre os *scores* nas amostras ComSW e SemSW, indicando que a remoção não influencia diretamente na avaliação da polaridade. Isso contradiz a expectativa de que a remoção das "palavras ditas vazias" poderia fazer diferença na análise do sentimento. Observou-se que a média do sentimento expresso nesta amostra de *tweets* é de positivo à neutro. Observou-se que os *scores* apresentam distribuição normal em torno da média, validando as análises estatísticas calculadas. E por fim com um fator de confiança de 95% pode-se inferir que não existe correlação entre a reputação (*rank*) e a polaridade das mensagens postadas. Sugerindo que o sentimento expresso nos *tweets* não influencia na inferência da reputação.

Como trabalhos futuros propõe-se estender a abordagem adotada para todos os 1232 usuários da listagem como forma de evidenciar os achados verificados nesta pesquisa. Além disso, pretende-se analisar a polaridade das mensagens em função da linha de tempo no qual as mensagens foram postadas. Pretende-se verificar a existência de mudanças na polaridade destas mensagens baseadas em algum evento do cotidiano.

REFERÊNCIAS

[1] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha, "Comparing and Combining Sentiment Analysis Methods," *Procs. of ACM COSN*, 2013.

[2] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, pp. 1-167, 2012.

[3] S. Bethard, H. Yu, A. Thornton, V. Hatzivassiloglou, and D. Jurafsky, "Automatic extraction of opinion propositions and their holders," in *2004 AAAI Spring Symposium on Exploring Attitude and Affect in Text*, 2004, p. 2224.

[4] S.-M. Kim and E. Hovy, "Automatic identification of pro and con reasons in online reviews," in

Proceedings of the COLING/ACL on Main conference poster sessions, 2006, pp. 483-490.

[5] A. Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums," *ACM Transactions on Information Systems (TOIS)*, vol. 26, p. 12, 2008.

[6] L. Weitzel, J. P. M. de Oliveira, and P. Quaresma, "Exploring Trust to Rank Reputation in Microblogging," in *Database and Expert Systems Applications*, 2013, pp. 434-441.

[7] E. d. B. A. Ferreira, "Análise de sentimento em redes sociais utilizando influência das palavras". Dissertaçao de mestrado, Universidade Federal de Pernambuco, Recife, 2010.

[8] N. R. Silva, D. Lima, and F. Barros, "SAPair: Um Processo de Análise de Sentimento no Nível de Característica," in *4nd International Workshop on Web and Text Intelligence (WTI'12), Curitiba*, 2012.

[9] R. L. Cilibrasi and P. M. Vitanyi, "The google similarity distance," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 19, pp. 370-383, 2007.

[10] P. Gonçalves, F. Benevenuto, and V. Almeida, "O Que Tweets Contendo Emoticons Podem Revelar Sobre Sentimentos Coletivos?," in *II Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2013), Maceió*, 2013.

[11] E. Boiy and M.-F. Moens, "A machine learning approach to sentiment analysis in multilingual Web texts," *Information retrieval*, vol. 12, pp. 526-558, 2009.

[12] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *Proceedings of LREC*, 2006, pp. 417-422.

[13] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, pp. 39-41, 1995.

[14] L. K. Wives and S. Loh, "Recuperação de informações usando a expansão semântica ea lógica difusa," in *Congreso Internacional En Ingenieria Informatica, ICIE*, 1998.