Uso de mineração de textos para a identificação de postagens com informações de localização

Silas F. Moreira[1], **Maruscia Baklizky**[1], **Luciano A. Digiampietri**[1]

[1]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (USP)

silas.moreira@usp.br, maruscia.baklizky@usp.br, digiampietri@usp.br

Abstract. *Information from social networks is very useful for different tasks. It is possible, for example, to quickly identify evidence of an epidemic, voting trends for an election, or user satisfaction with services or products. Several text mining related tasks that use information from social networks can take advantage of geographic information about the user's location. However, the majority of posts in social networks do not have an explicit reference to their geolocation. The goal of this paper is to compare the efficiency of text mining techniques to identify whether or not a post contains information about a location.*

Resumo. *As informações contidas nas redes sociais são cada vez mais úteis para diferentes tarefas. É possível, por exemplo, identificar de maneira bastante rápida indícios de uma epidemia, tendências de voto para uma eleição, ou a satisfação de usuários em relação a serviços ou produtos. Diversas tarefas de mineração de textos em redes sociais conseguem tirar vantagem de informações geográficas sobre a localização do usuário. Porém, a grande maioria das postagens em redes sociais não possui uma referência explícita à sua geolocalização. O objetivo deste trabalho é comparar a eficácia de algumas técnicas de mineração de textos para identificar se uma postagem contém ou não informações sobre uma localidade.*

1. Introdução

As redes sociais conquistaram uma alta popularidade atualmente, tornando-se um grande repositório de informações sobre seus usuários, por meio de seus perfis e de suas publicações. Por conta disso, a análise dos conteúdos provenientes dessas redes tem recebido cada vez mais atenção e tem possibilitado uma série de estudos envolvendo o comportamento de seus usuários.

Informações sobre a geolocalização de usuários são importantes para diversas tarefas, como recomendação de locais de interesse, de rotas e de produtos. Para algumas aplicações, como o traçado de rotas em tempo real, essa informação é fundamental. Por outro lado, outras aplicações, como sistemas de recomendação de produtos, podem ser mais efetivas com o uso da informação de localização, porém, podem funcionar adequadamente sem ela.

Com o uso de informações de redes sociais é possível, por exemplo, detectar de maneira antecipada locais com alta incidência de uma doença com base nas informações contidas nas trocas de mensagens. Porém, na maioria dessas redes, quando a informação de geolocalização não é requerida, observa-se que a frequência do uso de marcadores

precisos de localização (*geo-tags*) é bastante baixa. Assim, é interessante que sejam desenvolvidos sistemas computacionais que possam, com base nas informações postadas em redes sociais, inferir a geolocalização dessas postagens.

Nesse contexto, a mineração de texto apresenta métodos e ferramentas que podem auxiliar no processo de identificação de informações de localização [Berry and Castellanos 2004, Feldman and Sanger 2007, Manning et al. 2008, Aggarwal and Zhai 2012]. O presente artigo está focado na identificação se uma postagem possui ou não informação de localidade e nele foram comparadas diferentes estratégias, considerando-se as medidas de acurácia e precisão.

As três hipóteses que nortearam o desenvolvimento desta pesquisa são: (i) técnicas simples (baseadas apenas em frequência de palavras) podem alcançar resultados satisfatórios na indicação de que uma postagem possui ou não referência a uma localidade; (ii) algoritmos de classificação podem atingir resultados superiores aos das técnicas simples; (iii) a combinação dessas duas estratégias pode aumentar a acurácia dos resultados.

2. Materiais e Métodos

Neste trabalho foram utilizadas postagens públicas da rede social *Facebook*. Tais postagens não eram explicitamente geolocalizadas e foram manualmente classificadas com a indicação de contendo ou não alguma informação de localização. Foram consideradas informações de localização a citação explícita do nome de um determinado lugar, seja comercial (por exemplo, lojas e hotéis) ou político (por exemplo, cidades e países). Erros de grafia foram desconsiderados para a classificação, de forma que mesmo com a apresentação de erros de escrita nos nomes das localizações, a postagem ainda seria classificada como contendo informação de localização.

A tarefa completa de geolocalizar compreende gerar como resultado uma referência a uma localização no globo terrestre. Porém, o escopo do presente artigo restringe-se a identificar se uma determinada postagem contém ou não uma informação de localização, não tratando da geolocalização propriamente dita e dos problemas de ambiguidade relacionados.

Para testes e validação, foram utilizadas 187 postagens públicas de páginas do *Facebook*, sendo que 97 foram selecionadas da página "Loucos por Viagens", que apresenta informações e críticas sobre as diversas viagens de seus autores. As postagens públicas foram obtidas com o auxílio da API disponibilizada pela própria rede social e classificadas manualmente, indicando se contêm ou não alguma informação de localização.

A partir da classificação, utilizou-se a medida estatística TF-IDF [Jones 1972], que mede a frequência de cada palavra em uma postagem, em relação à sua frequência em todo o *corpus* de mensagens com ou sem informações de localidade. Com essas informações, tornou-se possível ordenar as mensagens de acordo com a soma das medidas de TF-IDF de cada palavra que compõe a mensagem. Neste trabalho, o cálculo da pontuação TF-IDF de uma postagem em relação a um *corpus* foi realizado de acordo com a equação 1, sendo t cada um dos termos pertencentes a postagem P, TF_t a frequência relativa do termo t na postagem P, e IDF_t o inverso da frequência relativa do termo t no *corpus* utilizado.

$$\sum_{t \in P} log\left(1 + TF_t * IDF_t\right) \tag{1}$$

A pontuação TF-IDF pode ser utilizada de diferentes formas. Duas pontuações para cada postagem foram calculadas e estas foram analisadas de três formas diferentes. Ambas as pontuações utilizaram a equação 1, porém, para uma o *corpus* utilizado foi formado apenas pelas postagens que contêm indicação de localidade, já para a outra, o *corpus* foi composto apenas por postagens que não continham indicação de localidade.

Com base nas duas pontuações atribuídas para cada postagem, três estratégias de classificação foram utilizadas. A primeira, denominada de *Max*, atribuiu para a respectiva postagem a classe que atingiu maior pontuação TF-IDF, ou seja, se o valor calculado para o *corpus* de mensagens que indicam localidade for maior, então a postagem será marcada como "indica localidade"; caso contrário, será classificada como "não indica localidade". A segunda estratégia, denominada de *Limiar*, utiliza um limiar considerando apenas a pontuação em relação ao *corpus* de mensagens que indicam localidade. Essa nota pode ser entendida como o pertencimento ou a adequação de uma postagem em relação às mensagens que indicam localidade. Para os experimentos, o limiar utilizado foi aquele que maximizou a acurácia. Por fim, analisou-se a relação entre a primeira pontuação obtida pela postagem e a soma das pontuações recebidas pela postagem. Esta estratégia foi denominada de *Relação* e a partir do valor da relação foi estabelecido um limiar para a classificação: são classificadas como "indicam localidade" as postagens cuja pontuação recebida no *corpus* que indica localidade dividida pela soma das duas pontuações foi maior do que um dado limiar.

Antes do computo das pontuações, foram realizadas combinações de diferentes estratégias de pré-processamento de forma a identificar as variações nos resultados pelo uso ou não destas estratégias. As estratégias utilizadas foram: remoção de *stop-words*, isto é, palavras que, a princípio, não agregam significado aos textos; e a radicalização (ou *stemming*). Adicionalmente, um filtro alternativo de palavras foi utilizado independente de um dicionário de *stop-words* ou de um algoritmo de radicalização: a exclusão de palavras pequenas (deixando apenas palavras com, pelo menos, um certo número de letras). O número mínimo de letras utilizado variou de 1 a 10.

A ideia de se medir a importância de palavras utilizando TF-IDF foi estendida para a importância de n-gramas. Assim, além do cálculo das pontuações para termos individuais (unigramas), também foram realizados os cálculos para bigramas e trigramas.

Uma abordagem alternativa para a representação das postagens, bastante utilizada na mineração de textos, é a de *bag-of-words*, na qual uma postagem é representada como um conjunto não ordenado de palavras. Com base nessa representação, diversas medidas de comparação podem ser calculadas entre diferentes postagens (por exemplo, a distância cosseno) ou outras estratégias, como o uso de classificadores, podem ser utilizadas. Esta representação foi utilizada como entrada para classificadores que tinham como objetivo identificar se uma postagem continha ou não uma informação de localidade. Assim, a classificação realizada foi binária. Para esta representação, duas estratégias de pré-processamento foram testadas: o uso ou não de radicalização e a seleção ou não das palavras (ou atributos) mais relevantes. Para esta abordagem optou-se pela utilização de um seletor de atributos, com o objetivo de identificar a importância relativa das palavras em relação às classes e selecionar o subconjunto "mais informativo" destes atributos. Foi utilizado um seletor baseado na correlação entre atributos, chamado *Correlation-based feature Subset Selection (CfsSubsetEval)* [Hall 2000].

Três classificadores foram utilizados: dois baseados no Teorema de Bayes (Rede Bayesiana - *Bayes Net* e *Naïve Bayes*) [John and Langley 1995] e um meta classificador que utiliza análise de componentes principais para a projeção dos atributos em um espaço no qual a variância está maximizada nas primeiras dimensões (*Rotation Forest*) [Rodriguez et al. 2006]. Todos os testes utilizaram validação cruzada com dez subconjuntos (*10-fold-cross-validation*). As implementações utilizadas do seletor de atributos e dos classificadores foram aquelas disponíveis no arcabouço Weka[1].

3. Resultados e Discussão

A combinação de diferentes estratégias de pré-processamento, com o uso de três abordagens diferentes para a classificação das postagens com base nas pontuações TF-IDF e o uso de unigramas, bigramas e trigramas produziu 360 resultados diferentes. A tabela 1 apresenta os resultados de acurácia para cada uma das abordagens e/ou combinações de estratégias de pré-processamento.

Tabela 1. Acurácia do uso de TF-IDF

		com radicalização									sem radicalização								
	Letras	Unigrama			Bigrama			Trigrama			Unigrama			Bigrama			Trigrama		
		Max	Limiar	Relação	Max	Limiar	Relação	Max	Limiar	Relação	Max	Limiar	Relação	Max	Limiar	Relação	Max	Limiar	Relação
sem remoção de stopwords	1	74.87%	83.96%	84.49%	81.82%	87.17%	83.96%	73.80%	74.87%	74.87%	72.73%	82.89%	84.49%	79.68%	86.63%	82.35%	75.40%	76.47%	76.47%
	2	75.40%	85.56%	84.49%	79.68%	82.35%	81.28%	70.05%	71.12%	70.59%	73.26%	85.03%	83.96%	81.28%	83.96%	81.82%	71.12%	72.19%	71.66%
	3	71.12%	84.49%	83.96%	79.68%	83.42%	82.35%	70.05%	70.05%	70.05%	74.33%	84.49%	80.75%	71.66%	73.80%	72.73%	64.71%	64.71%	64.71%
	4	76.47%	86.63%	85.56%	74.33%	75.94%	75.40%	66.84%	67.38%	67.38%	78.61%	86.63%	83.42%	69.52%	69.52%	69.52%	64.71%	64.71%	64.71%
	5	78.07%	85.56%	83.96%	73.26%	73.80%	73.80%	65.78%	65.78%	65.78%	85.03%	**87.70%**	85.03%	65.78%	65.78%	65.78%	64.71%	64.71%	64.71%
	6	83.42%	86.10%	82.35%	68.98%	69.52%	69.52%	65.78%	65.78%	65.78%	81.28%	84.49%	83.96%	66.84%	66.84%	66.84%	64.71%	64.71%	64.71%
	7	78.61%	82.89%	80.75%	66.84%	66.84%	66.84%	64.71%	64.71%	64.71%	77.01%	79.14%	78.07%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%
	8	79.68%	80.75%	80.75%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%	74.33%	75.40%	74.87%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%
	9	76.47%	77.01%	77.01%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%	70.59%	70.59%	70.59%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%
	10	72.19%	72.19%	72.19%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%	67.91%	67.91%	67.91%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%
com remoção de stopwords	1	80.75%	86.63%	84.49%	75.94%	78.07%	78.07%	67.91%	67.91%	67.91%	79.14%	85.03%	84.49%	80.21%	83.42%	83.42%	68.98%	68.98%	68.98%
	2	80.75%	87.17%	84.49%	76.47%	78.07%	78.07%	67.91%	67.91%	67.91%	78.61%	85.56%	84.49%	79.68%	81.82%	81.82%	70.05%	70.05%	70.05%
	3	80.75%	86.10%	83.42%	78.07%	79.68%	79.68%	67.91%	67.91%	67.91%	78.07%	85.03%	82.35%	71.66%	73.80%	72.73%	64.71%	64.71%	64.71%
	4	78.61%	87.17%	85.03%	72.19%	73.80%	73.26%	66.84%	67.38%	67.38%	78.07%	86.63%	83.42%	69.52%	69.52%	69.52%	64.71%	64.71%	64.71%
	5	81.82%	86.10%	83.96%	72.19%	72.73%	72.73%	65.78%	65.78%	65.78%	85.03%	**87.70%**	85.03%	65.78%	65.78%	65.78%	64.71%	64.71%	64.71%
	6	79.14%	85.03%	82.89%	68.98%	69.52%	69.52%	65.78%	65.78%	65.78%	81.28%	84.49%	83.96%	66.84%	66.84%	66.84%	64.71%	64.71%	64.71%
	7	77.01%	80.75%	79.14%	66.84%	66.84%	66.84%	64.71%	64.71%	64.71%	75.94%	79.14%	78.07%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%
	8	77.54%	78.61%	78.61%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%	74.33%	75.40%	74.87%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%
	9	74.33%	74.87%	74.87%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%	71.12%	71.12%	71.12%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%
	10	72.19%	72.19%	72.19%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%	68.45%	68.45%	68.45%	64.71%	64.71%	64.71%	64.71%	64.71%	64.71%

Duas células da tabela contêm o valor mais alto de acurácia observado (87,7%). Os dois resultados ocorreram com o uso de unigramas e utilizando como estratégia de classificação um limiar sobre o valor obtido pela pontuação TF-IDF em relação às postagens classificadas como "indicam localidade" nos quais o texto foi pré-processado por um processo de radicalização e excluindo-se palavras com menos de quatro letras. A remoção ou não das *stop-words* não alterou o valor deste resultado.

Em termos de características individuais, inicialmente destaca-se o uso de unigramas. Essa estratégia se sobressaiu ao uso de bigramas e trigramas, o que pode ter ocorrido devido ao *corpus* não ser muito grande, o que pode causar esparsidade entre as postagens no uso de bigramas e, principalmente, de trigramas. Em termos das três estratégias de classificação aplicadas (*Max*, *Limiar* e *Relação*), o uso do *Limiar* foi a estratégia que, na média, apresentou os melhores resultados de acurácia, seguido pela estratégia *Relação*. Uma vantagem do uso de *Limiar* é a não necessidade de formação de um *corpus* negativo (de postagens que não indicam localidade), precisando apenas de um *corpus* positivo e do estabelecimento de um limiar que pode ser adaptado para aumentar a precisão do sistema ou aumentar a revocação, de acordo com as necessidades do usuário.

[1]Weka 3: https://www.cs.waikato.ac.nz/ml/weka/ , acessado em 22/03/2018

O uso da radicalização, na média, apresentou resultados bastante parecidos com o não uso de radicalização (apesar de ter repercutido no melhor resultado de acurácia geral). A remoção de *stop-words*, na média, também aumentou a acurácia do sistema. Já os filtros de uso de palavras com apenas um número mínimo de letras teve seus melhores resultados para unigramas com tamanho entre quatro e seis letras.

A tabela 2 apresenta os resultados da classificação utilizando como base a representação *bag-of-words*. Destaca-se que a radicalização apresentou melhores resultados em todos os testes (em comparação ao não uso de radicalização) exceto em uma execução específica do *Rotation Forest*. Já o uso da seleção de atributos apresentou resultados melhores do que o uso de todas as palavras em todos os casos testados.

Tabela 2. Acurácia da classificação utilizando a representação *bag-of-words*.

	todas as palavras		palavras selecionadas	
Classificador	sem uso de radicalização	com uso de radicalização	sem uso de radicalização	com uso de radicalização
Bayes Net	82,89%	84,49%	87,70%	88,24%
Naive Bayes	80,21%	81,82%	83,96%	88,24%
Rotation Forest	81,82%	77,01%	79,14%	82,35%

O uso do classificador baseado em Rede Bayesiana (*Bayes Net*) apresentou resultados melhores ou iguais aos demais classificadores testados. Destaca-se ainda que os melhores resultados (88,24% de acurácia geral) atingidos pelos classificadores *Bayes Net* e *Naïve Bayes* com o uso de radicalização e seleção de atributos são levemente melhores do que os melhores resultados atingidos com o uso de TF-IDF (87,7%).

Por fim, avaliou-se a combinação do uso da pontuação TF-IDF e a representação *bag-of-words*. Optou-se por realizar testes apenas combinando os melhores resultados obtidos em cada uma das estratégias: no caso da pontuação TF-IDF, foi aplicada a radicalização, removidas as *stop-words*, e consideradas apenas palavras com cinco ou mais letras; já para a representação *bag-of-words*, considerou-se apenas o uso da radicalização. A tabela 3 apresenta os seis resultados obtidos pelos classificadores.

Tabela 3. Acurácia da classificação combinando a pontuação TF-IDF com a representação *bag-of-words*.

Classificador	todas as palavras	palavras selecionadas
Bayes Net	85,03%	89,84%
Naive Bayes	83,96%	86,10%
Rotation Forest	88,24%	89,30%

Observa-se que, na maioria dos casos, a combinação das características apresentou resultados melhores do que os resultados equivalentes das medidas individuais. Destaca-se que o melhor resultado entre todos os experimentos foi o obtido pelo classificador *Bayes Net* combinando a pontuação TF-IDF com a representação *bag-of-words*, utilizando seleção de atributos. A acurácia geral atingida foi de 89,84%.

4. Conclusões e Trabalhos Futuros

Neste trabalho foram avaliadas diferentes estratégias de mineração de textos para identificar se uma postagem em rede social online possui ou não informação de localidade,

usando como estudo de caso postagens veiculadas no *Facebook*. Duas principais abordagens foram avaliadas: o uso de uma pontuação dada às postagens com base em valores TF-IDF e o uso de classificadores para avaliar as postagens representadas como *bag-of-words*. Adicionalmente, a combinação destas duas abordagens também foi testada.

As três hipóteses de pesquisa foram confirmadas pelos resultados obtidos: (i) é possível obter resultados satisfatórios para o problema tratado considerando-se apenas o uso de estratégias simples baseadas na importância relativa das palavras; (ii) o uso de classificadores aumentou a acurácia da solução; e (iii) a combinação das duas estratégias anteriores proporcionou resultados ainda mais acurados.

Destaca-se que o uso de um limiar sobre a pontuação TF-IDF a partir do qual as postagens são classificadas como "indicam localidade" além de ter atingido uma acurácia satisfatória ainda permite uma variação de valores por parte do usuário a fim de maximizar a precisão na identificação dos elementos positivos ou a revocação dos elementos positivos (de acordo com as necessidades da aplicação).

Como trabalhos futuros, pretende-se aplicar as mesmas abordagens testadas neste artigo a um conjunto de dados oriundo de outras redes sociais online de forma a verificar a variabilidade dos resultados em diferentes contextos.

Agradecimentos

Este trabalho foi parcialmente financiado pelo Programa de Educação Tutorial (PET) do Ministério da Educação, pela CAPES e pelo CNPq.

Referências

Aggarwal, C. C. and Zhai, C. (2012). *Mining text data*. Springer Science & Business Media.

Berry, M. W. and Castellanos, M. (2004). Survey of text mining. *Computing Reviews*, 45(9):548.

Feldman, R. and Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.

Hall, M. A. (2000). Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 359–366, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

John, G. H. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 338–345.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Rodriguez, J. J., Kuncheva, L. I., and Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(10):1619–1630.