Análise de Sentimento em Documentos de Texto Financeiros com Múltiplas Entidades

Javier Zambrano Ferreira
Instituto de Computação
Universidade Federal do Amazonas
Manaus, Amazonas - Brasil
Email: zambrano.ferreira@gmail.com

***Resumo*—O volume de notícias publicados na Internet tornar-se impossível a análise manual de cada notícia. Com isso, ferramentas de análise de sentimento ou polaridade são utilizados para classificar os documentos em positivo, negativo e neutro. Uma abordagem comum dessas ferramentas é classificar o texto apenas referenciando uma entidade, entretanto na sua grande maioria mais de uma entidade é citada no texto. Um domínio de notícia em que há comparações constantes entre entidades é o domínio financeiro uma vez que a comparação entre empresas e seus desempenhos são confrontadas. Neste trabalho busca-se verificar a polaridade de cada entidade nos textos de domínio financeiro e por fim, avaliar a polaridade do documento. Para isso, utilizamos três etapas: 1) extração das entidades no texto; 2) os fragmentos de textos em que cada entidade é referenciada e 3) o cálculo da polaridade para cada entidade apenas sobre os textos na qual ela é citada.**

I. Introdução

O volume de informação disponível na Internet, seja em *Websites*, fóruns e página de notícias, é tão grande que é impossível a análise manual visando identificar o conteúdo relevante a um determinado domínio e a natureza deste conteúdo. Um tipo de análise de interesse desse conteúdo consiste em determinar a polaridade da opnião do autor em relação ao assunto em discussão, o que chamamos de análise de sentimento ou polaridade. Um exemplo de análise de sentimento é inferir se, em um texto sobre um produto, o autor do texto emite uma opnião favorável, neutra ou desfavorável em relação ao produto.

A análise de sentimentos tem sido usada em uma varidade de domínios de aplicação. Por exemplo, ela é útil para inferir automaticamente a opnião de um cliente sobre um certo produto de uma loja virtual com base em um comentário que este postou, a opinião de uma pessoa sobre um item postado em uma rede social, etc. Enquanto algumas técnicas gerais podem ser usadas para qualquer domínio a simples transposição do que vale em um domínio para o outro pode não ser bem sucedida. Como observado por [10], diversos termos pré-classificados como positivos em um domínio possuem uma conotação neutra em diferentes contextos.

Um domínio de particular interesse, e foco deste trabalho, é o domínio dos documentos financeiros. O interesse neste domínio se deve à hipótese de que notícias de caráter positivo ou negativo, relacionadas com uma companhia, podem afetar o desempenho financeiro desta companhia na bolsas de valores [1], [3], [4]. Assim, a polaridade de um documento de natureza financeira poderia ser usada para ajudar a prever tendências relacionadas com o desempenho de uma companhia.

Em termos de desenvolvimento de técnicas e algoritmos, o desafio, no caso de documentos financeiros, é maior, uma vez que, ao contrário de domínios como filmes e produtos, os autores dos textos não os avaliam por meio de notas [1]. Outra característica dos trabalhos neste domínio é a premissa de que os documentos são a respeito de uma única entidade [1], [8]. Em nosso trabalho contudo não partimos dessa premissa, uma vez que diversos documentos citam duas ou mais entidades, como o exemplo do texto dado na Figura 1.

> The ANALYSIS: The verdict against Samsung could give Nokia an edge. The Windows Phone is substantially different from Apple's iPhone operating system and hasn't landed in its legal sights, and some Wall Street analysts say that the verdict against Samsung is likely to slow growth of smartphones that run on Android.

Figura 1. Exemplo de texto com múltiplas entidades.

Como podemos observar na Figura 1, três entidades são citadas no documento: *Nokia*, *Apple* e *Samsung*. A abordagem sugerida em trabalhos anteriores [1], [3], infere a polaridade do texto como um todo para apenas uma entidade pré-determinada (por exemplo, por meio de uma consulta). Neste caso, contudo, a polaridade do texto é diferente para cada entidade, uma vez que o texto refere-se que a Nokia pode ganhar com a perda da Samsung no jugamento com a Apple. Nota-se que a entidade Apple é apenas citada como fabricante do iPhone e não por ter ganho o julgamento contra a Samsung, com isso sua polaridade é neutra. Porém, a entidade Nokia possui a polaridade positiva e a entidade Samsung é citada negativamente no texto. Assim, neste trabalho, o foco é identificar que entidades estão presentes no texto, quais fragmentos são relacionados a cada entidade e qual a polaridade em relação a cada entidade.

II. Trabalhos Relacionados

A análise de sentimentos tem sido empregada em diversos domínios, como resenhas de filmes e produtos. O primeiro

trabalho nesta linha foi proposto por [6], que demonstrou que a classificação de documentos de acordo com a sua polaridade é similar à classificação com base em seus tópicos. Os resultados apresentados demonstram que a classificação dos documentos por sua polaridade é tão desafiante quanto por tópicos, uma vez que é comum o uso de ironia e o contexto é importante para a definição da polaridade, já que palavras de cunho positivo podem ocorrer em frases negativas (o contexto) ou vice-versa.

Para lidar com o contexto da contextualização [10] propuseram uma abordagem em que explora características de frases para analisar o sentimento dos textos. O principal resultado demonstrado é que um conjunto léxico pré-classificado como positivo e negativo a priori não funciona sempre, pois depende do contexto em que o termo é utilizado. Os autores desse trabalho também demonstraram que termos classificados como positivos negativos são comuns em frases neutras.

Diferente dos dois trabalhos citados [12] demonstrou que a polaridade de um documento deve ser obtida com base em diferentes tópicos dentro do mesmo texto. Este trabalho é particularmente interessante para nós, uma vez que também consideramos que a polaridade deva ser tomada a partir de segmentos de texto. Os autores em [1] propuseram o uso de análise de polaridade do domínio financeiro, motivado pela possibilidade de previsão das reações do mercado de ações. O autor concluí que é possível aprender os termos mais usados e de maior impacto para medir a polaridade, com desempenho tão bom quanto o de avaliadores humanos. Eles também observaram que os modelos aprendidos no domínio financeiro não foram úteis quando aplicados a outros domínios.

Em [2], os autores estudaram a relação entre polaridade de textos associados a companhias com o seu desempenho no mercado de ações. Ao estudar um grande volume de dados do Twitter, os autores observaram que mudanças no estado emocional do público tem impacto dias depois no mercado financeiro.

Outros trabalhos que exploraram a relação entre polaridade e desempenho no mercado de ações foram propostos por [3], [4]. Estes trabalhos se basearam na categorização das emoções básicas do homem, segundo Darwin: raiva, medo e tristeza, entre outros. Também delimitaram os sentimentos de acordo com múltiplas dimensões ao invés de categorias discretas, por exemplo, o texto pode ser classificado como muito bom e não apenas positivo. Em suma, sua abordagem baseia-se no uso de um conjunto léxico com termos positivos e negativos, com o apoio da teoria de Darwin que tenta identificar quão intenso é esse sentimento. Para os termos positivos houve uma alta revocação, porém uma baixa precisão. Quanto aos termos negativos, ocorreu o inverso. Por fim, o autor conclui que o mapeamento direto dos termos do texto com os termos da teoria de Darwin não é algo simples de ser feito.

Finalmente, [11] também apresentam um estudo sobre o impacto da polaridade na previsão financeira. Os autores analisam notícias financeiras com base em diferentes representações textuais. Neste trabalho, os autores observaram que há uma correlação entre o preço futuro de uma ação e os seus preços tomados no momento em que os artigos sobre ela são publicados, quando considerados em conjunto com as polaridades dos seus termos presentes nestes artigos.

A abordagem utilizada neste trabalho é analisar a polaridade para cada entidade citada no texto.

III. Implementação da Proposta

O trabalho está sendo implementado em três etapas: 1) criação da coleção para os experimentos; 2) implementação e uso de ferramentas para extração de entidades, extração de fragmentos de textos em que cada entidade é citada e por fim, o cálculo da polaridade; 3) avaliação dos resultados. A Figura 2 mostra as etapas da implementação proposta

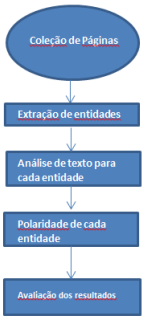

Figura 2. Etapas para implementação da proposta.

Um grande desafio deste trabalho consiste em obter uma coleção de páginas Web que possui apenas notícias do domínio financeiro, porém sem o comportamento do mercado de ações. Com isso, a primeira etapa do projeto consistiu na criação dessa base por meio da coleta de documentos financeiros de diferentes sites: *Bloomberg*, *New York Times*, *Financial Times*, *Reuters* e *Wall Street Journal*. A base consiste em mais de 20.000 páginas coletadas, entre várias entidades de diferentes ramos de negócios. Para poder determinar, qual a proporção de documentos em que a polaridade difere entre entidades no texto, 1000 páginas foram escolhidas aleatoriamente e cinco entidades foram utilizadas: *Apple*, *Microsoft*, *Google*, *Nokia* e *Samsung*. A diminuição no escopo de entidades permite verificar a relação entre elas. A Tabela I mostra o número de entidades para o conjunto de 1000 páginas.

Tabela I
Número de documentos em que a entidade é citada.

Entidade	Número de documentos
Nokia	58
Google	398
Microsoft	254
Apple	958
Samsung	313

Essas páginas formam a base de treino desse trabalho uma vez que foram e estão sendo avalidas por seres humanos ao longo do processo. Para isso, um sistema de avaliação foi desenvolvido [5], a página com a notícia é apresentado ao usuário e a opção de entidades citadas no texto, dentro do

conjunto das cinco entidades selecionadas. A Figura 3, mostra a interface do sistema.

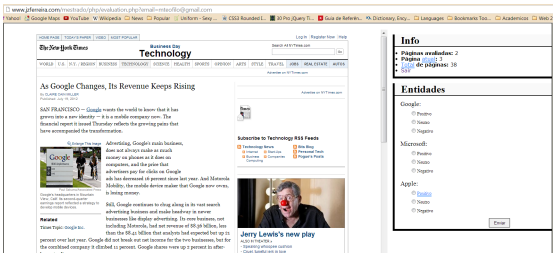

Figura 3. Interface do Sistema de Avaliação.

Para cada avaliador um conjunto de 40 páginas são avaliadas sendo que cada entidade contida no texto é avaliada entre positivo, negativo e neutro. A Tabela II apresenta a distribuição da polaridade para cada entidade no conjunto de 503 páginas já avaliadas.

Tabela II
DISTRIBUIÇÃO DE POLARIDADE POR ENTIDADE.

Entidade	Pos	Neg	Neutro
Nokia	12	13	9
Google	66	94	20
Microsoft	22	57	14
Apple	148	267	87
Samsung	45	64	35

Para segunda etapa, extração de entidades, foi utilizada a ferramenta Illinois Named Entity Tagger [7], o estado da arte para extração de entidades. A entrada para a ferramenta são textos em formato TXT e sua saída são as entidades marcadas no próprio documento. É possível definir o nível de marcação das entidades sendo as quatro classes clássicas: pessoas, organizações, localizações e outros como padrões. Porém, este trabalho possui interesse apenas nas classes pessoas e organizações.

A segunda parte desta etapa consiste no uso de ferramentas e algoritmos para a extração de fragmentos de textos em que cada entidade é referenciada no documento. Uma ferramenta que representa o estado da arte na identificação desses fragmentos e as entidades por eles referenciadas é a Beautiful Anaphora Resolution Toolkit (BART). O BART [9] utiliza-se de referência anafóras para verificar no documento em que trechos tal entidade é citada.

O próximo passo deste trabalho, consiste em finalizar a base de treino em 1000 páginas e iniciar a etapa de cálculo de polaridade para as entidades, apenas com os fragmentos de textos em que são citadas, com técnicas de máquinas de aprendizagem.

IV. CONSIDERAÇÕES FINAIS

As etapas implementadas neste trabalho demonstraram que a abordagem de verificar a polaridade de cada entidade em um texto é desafiador. A obtenção de uma base de dados em que diversas entidades citadas em texto são classificadas individualmente não foi encontrada na literatura. Para prosseguir o trabalho, a construção da base de treino com documentos avaliados em positivo, negativo e neutro por seres humanos consumiu o maior tempo deste trabalho. A Tabela 2 permite inferir que as notícias negativas são as que influenciam mais, uma vez que há uma grande disparidade entre as classes. O maior desafio é construir a base de dados com as páginas classificadas por seres humanos.

O próximo passo é propor algoritmos e implementar métodos da literatura com técnicas de máquinas de aprendizagem para o cálculo de polaridade para cada entidade.

REFERÊNCIAS

[1] P. Azar. *Sentiment Analysis Financial News*. PhD thesis, Hardvard College, 2009.

[2] J. Bollen, H. Mao, and X.-J. Zeng. Twitter mood predicts the stock market. 1(2):1–8, 2010.

[3] A. Devitt and K. Ahmad. A lexicon for polarity: Affective content in financial news text. *Proceedings of Language For Special Purposes*, 2007.

[4] A. Devitt and K. Ahmad. Sentiment polarity identification in financial news: A cohesion-based approach. *45th Annual Meeting of the Association for Computational Linguistics*, June 2007.

[5] J. Z. Ferreira. Sistema de avaliação. 2012.

[6] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP '02 Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, 2002.

[7] L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *Conference on Natural Language Learning*, 2009.

[8] R. P. Schumaker and H. Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Trans. Inf. Syst.*, 27:12:1–12:19, March 2009.

[9] Y. Versley, M. Poesio, and X. Yang. Conference systems based on kernel methods. In *Proceedings of the 22nd International Conference on Computational Linguistics*, 2008.

[10] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT '05 Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005.

[11] I. H. Witten, E. Frank, and M. A. Hall. *Data mining : practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, CA, USA, 3rd edition, 2011.

[12] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 427 – 434, November 2003.