PODER EXECUTIVO
MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO AMAZONAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Proposta de Dissertação de Mestrado

Análise de sentimento em documentos de texto financeiros com multíplas entidades

Aluno:
Javier Zambrano Ferreira

Orientador:
Prof. Dr. Marco Cristo

28 de fevereiro de 2012

1 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 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 opinião 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 opinião favorável, neutra ou desfavorável em relação ao produto.

A análise de sentimentos tem sido usada em uma variedade de domínios de aplicação. Por exemplo, ela é útil para inferir automaticamente a opinião de um revisor a partir da resenha que ele escreveu sobre um filme, a opinião 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 [e Janyce Wiebe e Paul Hoffmann, 2005], 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 bolsa de valores [Azar, 2009, e Huina Mao e Xiao-Jun Zeng, 2010, e Khurshid Ahmad, 2007a]. 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 [Azar, 2009]. Outra característica dos trabalhos neste domínio é a premissa de que os documentos são a respeito de uma única entidade [Azar, 2009, e Huina Mao e Xiao-Jun Zeng, 2010, Schumaker, 2009]. Em nosso trabalho, contudo, não partimos desta premissa, uma vez que diversos documentos citam duas ou mais entidades, como o exemplo dado na Tabela 1.

Como podemos observar na Tabela 1, duas entidades são citadas no documento: *Amazon* e *Apple*. Além disso, note que a polaridade para cada entidade citada é distinta. A abordagem sugerida em trabalhos anteriores

Amazon's new Kindle Fire was a hot item during the holiday shopping season, and one analyst believes the new Amazon tablet may have cost Apple well over $1 billion in holiday iPad sales. Morgan Keegan analyst Travis McCourt on Tuesday lowered his December-quarter iPad sales estimate from 16 million units to 13 million. Hot sales of the Kindle Fire ahead of the holidays are responsible for trimming sales of Apple's iPad by between 1 million and 2 million units, the analyst believes, making Amazon's new slate the main reason for McCourt's slashed forecast. On the low end of McCourt's estimate, the Kindle Fire cost Apple at least $500 million considering the iPad 2's $500 entry-level price point. If the Kindle Fire was indeed responsible for cutting iPad sales by 2 million units, Amazon tablet sales cost Apple a minimum of $1 billion. Considering the range of available iPad 2 models that sell for between $500 and $830 each, however, that figure would likely be significantly higher. Amazon announced last week that it sold more than 4 million Kindles during the holiday shopping season, noting that the Kindle Fire was its most popular device. McCourt believes total Kindle Fire sales for the 2011 holiday shopping season were between 4 million and 4.5 million units.

Tabela 1: Exemplo de documento com múltiplas entidades

[Azar, 2009, e Khurshid Ahmad, 2007b], 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 é positiva para Amazon e negativa para Apple. Logo, acreditamos que, para uma correta análise de sentimento, o texto todo não deve ser tratado com uma única polaridade. Ele deveria ser considerado uma combinação de fragmentos relacionados com diferentes entidades e, possivelmente, diferentes polaridades.

Assim, neste trabalho de mestrado, o foco é de identificar que entidades estão presentes no texto, quais fragmentos são relacionados a cada entidade e qual a polaridade em relação a cada entidade.

O restante da proposta está dividido da seguinte maneira. Na seção 2, são apresentados os objetivos. Na seção 3, descrevemos os trabalhos relacionados. Na seção 4, descrevemos a metodologia a ser empregada no trabalho. Na seção 5, apresentamos alguns resultados preliminares. Na seção 6, mostramos nosso cronograma de atividades.

2 Objetivos

2.1 Geral

O objetivo deste trabalho é analisar documentos textuais com múltiplas entidades, detectando a polaridade do documento em relação a cada entidade particular.

2.2 Específicos

Este trabalho possui os seguintes objetivos específicos:

1. Fazer pesquisa bibliográfica envolvendo análise de sentimentos, em particular, no domínio financeiro;
2. Criar uma coleção de documentos financeiros para análise de polaridade, com múltiplas entidades. Para esta coleção, verificar a distribução das entidades e, para uma amostra dela, rotulá-las de acordo com as polaridades observadas, usando avaliadores humanos;
3. Implementar método da literatura que será usado como base de comparação;
4. Sugerir e implementar métodos para detectar entidades no texto;
5. Sugerir e implementar métodos para determinar trechos dos textos associados a cada entidade;
6. Sugerir e implementar métodos para determinar a polaridades desses trechos;
7. Avaliar método proposto, comparando-o com método base de comparação.

3 Revisão Bibliográfica

A análise de sentimentos tem sido empregada em diversos domínios, como resenhas de filmes e produtos. O primeiro trabalho nesta linha foi proposto por [e Lillian Lee e Shivakumar Vaithyanathan, 2002], que demonstrou que a classificação de documentos de acordo com a sua polaridade é similar à classicação com base em seus tópicos. Os experimentos foram realizados com base nas resenhas de filmes do site *Internet Movie Database* (IMdB), em que apenas os documentos que continham uma nota associada ao texto

do documento foram utilizados. Nesse trabalho, os autores usaram os classificadores Naive Bayes, Máxima Entrópia e Support Vector Machine (SVM) [Witten, 2011]. Destes, o Naive Bayes apresentou o pior desempenho. Os demais classificadores tiveram desempenho comparável ao realizado por pessoas. 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 frases negativas (o contexto) ou vice-versa.

Para lidar com o problema de contextualização, [e Janyce Wiebe e Paul Hoffmann, 2005] propuseram uma abordagem em que explora características de frases para analisar o sentimento dos textos. A base de dados usada foi *Multi-perspective Question Answering* (MPQA) cujo conteúdo consiste de documentos em língua inglesa, com conteúdo detalhado e forte significado emocional. 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. Por exemplo, no segmento de texto na Tabela 2, o termo Trust expressa um sentimento positivo. Porém, no contexto dado, a palavra não é usada para expressar um sentimento e sim, o título da entidade. Os autores desse trabalho também observaram que termos classificados como positivos e negativos são comuns em frases neutras.

Philip Clapp, president of the National Environment Trust, sums up well the general thrust of the reaction of environmental movements: "There is no reason at all to believe that the polluters are suddenly going to become reasonable

Tabela 2: Exemplo de termo positivo usado em contexto do nome de uma entidade

Diferente dos dois trabalhos citados, [Yi, 2003] demonstraram 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.

O autor em [Azar, 2009] foi o primeiro a propor o uso de análise de polaridade no domínio financeiro, motivado pela possibilidade de previsão das reações do mercado de ações. Eles usaram como base de dados a *Reuters Key Developments Corpus*, que contém notícias no período de 1998 a 2009. Das companhias nesta coleção, o autor utilizou somente as com mais de 20 notícias. Seus resultados foram obtidos com o uso de técnicas de processamento de linguagem natural e classificadores baseados em Árvores de Decisão

e SVM. O autor conclui que é possível aprender os termos mais usados e de maior impacto para medir polaridade, com desempenho tão bom quanto o de avaliadores humanos. Ele também observou que os modelos aprendidos no domínio financeiro não foram úteis quando aplicados a outros domínios.

Em [e Huina Mao e Xiao-Jun Zeng, 2010], os autores estudaram a relação entre a polaridade de textos associados a companhias com o seu desempenho no mercado de ações. O *Twitter* foi usado como base para experimentação. 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. Seus resultados foram alcançados com simples técnicas de processamento de linguagem natural.

Outros trabalhos que exploraram a relação entre polaridade e desempenho no mercado de ações foram propostos por [e Khurshid Ahmad, 2007a] e [e Khurshid Ahmad, 2007b]. Estes trabalhos se basearam na categorização das emoções básicas do homem, segundo Darwin: raiva, medo e tristeza, entre outras. Também delimitaram os sentimentos de acordo com múltiplas dimensões ao invés de categorias discretas. Duas dimensões primárias foram utilizadas: um eixo bom-mal e outro eixo de forte-fraco. Os experimentos foram realizados com bases em notícias e comportamento do mercado de ações relativos a duas companhias aéreas da Irlanda. A técnica para mensurar a polaridade do texto consistiu em construir um grafo que representa o texto todo. Neste grafo, os nós são termos do texto com seus respectivos valores de polaridade (obtidos com a ferramenta SentiWordNet, que mede os termos em positivos e negativos de acordo com a WordNet). 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, [Schumaker, 2009] também apresentam um estudo sobre o impacto da polaridade na previsão financeira. Os autores analisam notícias financeiras com base em diferentes reprsentações textuais: *bag of words*, sintagmas nominais e nome de entidades. 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 artigos sobre ela são publicados, quando considerados em conjunto com as polaridades dos seus termos presentes nestes artigos.

O trabalho que estamos propondo nessa dissertação é distinto dos trabalhos citados nesta seção uma vez que realizamos a análise de sentimentos com base em várias entidades.

4 Metodologia

O trabalho será implementado em três etapas, descritas nos parágrafos a seguir.

A primeira etapa é a criação de uma coleção para os experimentos. Como fonte de dados, usaremos a Reuters Key Developments Corpus, uma base que contém notícias sobre mercados de ações. Outra possível fonte de dados é a Multi-perspective Question Answering (MPQA), uma base rica em termos com forte carga de emoção. Porém, é necessário filtrá-la uma vez que as suas notícias são de vários domínios e não apenas financeiro, o nosso interesse neste trabalho. O último passo para a criação da base é coletar documentos financeiros de diferentes *sites*: Bloomberg, Business Wire, CNN Money, Wall Street Journal, Financial Times, Forbes e Reuters. Nesta fase ainda analisaremos a distribução de entidades nos documentos e, para uma amostra de documento, iremos verificar a sua polaridade em relação às suas diversas entidades.

A segunda etapa consiste em implementar e utilizar métodos da literatura para a detecção de entidades em documentos textuais. Uma abordagem simples a ser utilizada é verificar se o termo é uma entidade na Wikipedia. O uso de técnicas de processamento de linguagens naturais na extração e reconhecimento de nome de entidades (pessoas ou empresas) é o foco deste trabalho nessa etapa. Além disso, serão usados algoritmos para a extração de terminologias. Com isso poderemos verificar os termos relevantes no domínio financeiro e usar no cálculo da polaridade. Os algoritmos a serem utilizados neste trabalho são Naive Bayes, Árvores de decisão e SVM [Witten, 2011]. A identificação dos termos permite também a criação de uma gramática, permitindo a criação de extratores. A identificação dos termos usa métodos de processamento de linguagens como parte-do-discurso, reconhecimento na variação de termos e o uso da freqüência de termos (tf e idf).

A segunda parte dessa 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 utiliza-se de referência anafóras para verificar no documento em que trechos tal entidade é citada. Além do BART, experimentaremos com estratégias mais simples como (a) considerar que todo o parágrafo em que uma entidade aparece se refere a ela ou (b) considerar que um fragmento de texto se refere à última entidade citada.

Após a extração dos fragmentos de texto em que uma entidade aparece, o cálculo da polaridade será feito, não em todo documento, mas em cima

desses fragmentos e tendo como saída a polaridade do termo no texto. Para tanto, iremos usar métodos de aprendizado de máquina como Naive Bayes e SVM.

A última etapa do nosso trabalho é a avaliação dos resultados. O uso de uma coleção com documentos pré-classificados, pelo menos como positivos e negativos, será usada no treino. A avaliação consistirá na avaliação da precisão, revocação e acurácia dos resultados. Por fim,o desempenho do sistema também será usado como parâmetro de avaliação. Duas medições serão feitas: 1) tempo de execução e 2) tempo de execução x custo.

5 Resultados Preliminares

Atualmente, estamos em nossa primeira etapa que consiste em criar a coleção de estudos e verificar, nesta coleção, sua distribução de entidades e polaridade por entidade. Com isso, esperamos validar algumas idéias básicas da nossa proposta.

Assim, dois experimentos foram realizados. Para tais experimentos, a coleção utilizada é formada por páginas da *Bloomberg* catalogadas na seção financeira com o total de 3831 páginas. Tal coleção foi criada com o uso de um coletor de páginas. O primeiro experimento consiste em verificar o número de entidades em cada documento. A técnica usada verifica se os termos contidos nos textos representam uma das companhias da lista *Forbes 2000*, que contém as 2000 maiores companhias globais. Caso o termo esteja, contabiliza-se uma entidade no documento.

A Tabela 3 mostra o total de documentos com o número de entidades citadas no texto. Nesta tabela observamos que o número mais comum de entidades por documento é cinco. Contudo, note que isso não significa que o texto contenha comparações entre tais entidades. Há documentos sobre o mercado financeiro como um todo, em que várias entidades, de diferentes ramos de negócios, são descritas com relação ao comportamento de suas ações financeiras.

Na Tabela 4, temos um documento que descreve as ações financeiras de várias entidades, a saber,*Google*, *IBM*, *Intuitive Surgical Inc.*, *JDS Uniphase Corp.*, *Marathon Petroleum* e *Microsoft*. O texto é sobre a queda nas ações do Google e a alta nas da Microsoft, além de descrever outras entidades de ramos distintos. Observamos que este documento é negativo em relação ao Google e positivo em relação à Microsoft. Uma questão interessante é determinar qual a proporção de documentos em que estas diferentes polaridades se observam.

Assim, nosso segundo experimento consiste em verificar a polaridade para um conjunto de entidades em diversos documentos. Para isso, cinco entidades

Número de Entidades	Total de Documentos
1	9
2	214
3	483
4	650
5	730
6	571
7	427
8	268
9	182
10	120
11	84
12	33
13	18
14	10
≥ 15	32

Tabela 3: Documentos com o número de entidades citadas em seu texto

foram utilizadas: Apple, Microsoft, Google, Nokia e Samsung. Tais entidades representam empresas de um mesmo ramo de negócio: celulares. Em seguida, selecionamos um conjunto de cem páginas em que pelo menos duas destas entidades são citadas. A polaridade foi calculada manualmente. Note que, neste experimento, apenas duas classes de polaridade foram levados em conta: positivo e negativo.

Do conjunto de documentos examinados, notamos que 28% apresentam uma única polaridade, sendo 16% positivos e 12% negativos. Todos os demais documentos (72% da amostra) apresentam as polaridades positiva e negativa. Estes resultados confirmam que diferentes polaridades podem ser observadas em um mesmo documento.

As Tabelas 5 e 6 mostram a polaridade das entidades Apple e Microsoft, respectivamente, em relação às demais entidades. Como estamos considerando pares de entidades, são quatro as possíveis classificações: positivo x positivo, positivo x negativo, negativo x positivo e negativo x negativo. A maior relação positivo x positivo foi Microsoft e Nokia, com sete documentos, entre as 100 páginas, que citam positivamente ambas as entidades. Isso é um reflexo da estreita relação entre as duas empresas. A relação positivo x negativo é maior entre Apple e Microsoft, algo esperado se notarmos que tais entidades são concorrentes em seus setores de atuação. Este experimento

Google Inc. (GOOG) fell 8.4 percent to $585.99, after losing 8.4 percent for the biggest loss in the Standard & Poor's 500 Index. The owner of the world's most popular Internet search engine reported fourth-quarter revenue and profit that missed analysts' estimates as an economic slowdown in Europe crimped international sales. Insmed Inc. (INSM) surged 32 percent, the biggest gain in the Russell 2000 Index, to $5.01. The Food and Drug Administration lifted a suspension of clinical trials of the company's experimental drug Arikace for patients with non- tuberculous mycobacteria lung disease.
International Business Machines Corp. (IBM) advanced 4.4 percent, the most since July 19, to $188.52. The world's biggest computer-services provider forecast 2012 earnings exceeding analysts' estimates after fourth-quarter profit rose 4.4 percent because of rising software demand.
Intuitive Surgical Inc. (ISRG) sank 6.1 percent, the most since Aug. 8, to $445.68. The maker of a robotic system to perform surgery said annual growth in its da Vinci surgical procedures slowed to 27 percent in the fourth quarter from 30 percent in the third quarter.
JDS Uniphase Corp. (JDSU) rose 4 percent to $13.45, the highest price since Sept. 15. The maker of fiber-optic equipment was raised to "buy" from "hold" at Stifel Nicolaus & Co., which said the company has a "sustainable competitive advantage."
Marathon Petroleum Corp. (MPC) (MPC US) jumped 3.7 percent to $37.17, the highest price since Nov. 11. The largest independent U.S. refiner by market value outperformed competitors after hedge fund Jana Partners LLC bought a 5.5 percent stake in the company.
Microsoft Corp. (MSFT) climbed 5.7 percent, the biggest gain in the Dow Jones Industrial Average, to $29.71. The world's largest software maker reported second-quarter profit that beat estimates, lifted by holiday sales of Xbox machines and Kinect sensors, as well as corporate software demand.

Tabela 4: Exemplo de documento com descrição das ações no mercado financeiro

sugere que um conhecimento prévio das relações entre as entidades pode ser útil para calcular suas polaridades.

		Microsoft		Google		Nokia		Samsung	
		POS	NEG	POS	NEG	POS	NEG	POS	NEG
Apple	POS	8	12	3	0	6	11	4	2
	NEG	6	1	4	2	1	0	4	1

Tabela 5: POS: positivo e NEG: Negativo. Entidade Apple com relação as demais

		Apple		Google		Nokia		Samsung	
		POS	NEG	POS	NEG	POS	NEG	POS	NEG
Microsoft	POS	8	6	3	6	7	2	3	2
	NEG	12	1	3	4	0	1	1	3

Tabela 6: POS: positivo e NEG: Negativo. Entidade Microsoft com relação as demais

6 Cronograma

Atividades	Ano de 2012										Ano de 2013		
	Mar	**Abr**	**Mai**	**Jun**	**Jul**	**Ago**	**Set**	**Out**	**Nov**	**Dez**	**Jan**	**Fev**	**Mar**
Créditos e Revisão bibliográfica	**X**	**X**											
Criação de uma base de documentos	**X**	**X**											
Definição e implementação de modelos e métricas de avaliação			**X**	**X**	**X**	**X**	**X**						
Experimentação e avaliação dos resultados						**X**	**X**	**X**	**X**				
Escrita de Artigo e da Dissertação									**X**	**X**	**X**	**X**	**X**

Referências

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

[e Huina Mao e Xiao-Jun Zeng, 2010] e Huina Mao e Xiao-Jun Zeng, J. B. (2010). Twitter mood predicts the stock market. 1(2):1–8.

[e Janyce Wiebe e Paul Hoffmann, 2005] e Janyce Wiebe e Paul Hoffmann, T. W. (2005). 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.*

[e Khurshid Ahmad, 2007a] e Khurshid Ahmad, A. D. (2007a). A lexicon for polarity: Affective content in financial news text. *Proceedings of Language For Special Purposes.*

[e Khurshid Ahmad, 2007b] e Khurshid Ahmad, A. D. (2007b). Sentiment polarity identification in financial news: A cohesion-based approach. *45th Annual Meeting of the Association for Computational Linguistics.*

[e Lillian Lee e Shivakumar Vaithyanathan, 2002] e Lillian Lee e Shivakumar Vaithyanathan, B. P. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP '02 Proceedings of the ACL-02 conference on Empirical methods in natural language processing.*

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

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

[Yi, 2003] Yi, J. e Nasukawa, T. e. B. R. e. N. W. (2003). 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.