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

24 de fevereiro de 2012

1 Introdução

O volume de informação disponível na Internet seja em *Websites*, fóruns e página de nóticias, é tão grande que é praticamente impossível a analíse manual visando idenfiticar algum conteúdo relavante a um determinado domínio. Um tipo de analíse desse conteúdo verifica o assunto do texto em discussão. Visando otimizar essa analíse e apresentar uma resposta ao interessado, usam-se ferramentas de análise de sentimentos, também conhecida como ferramentas de polaridade, as ferramentas aferem se o texto sobre um tópico é positivo, negativo ou neutro.

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 de que vale em um domínio para o outro pode não ser bem sucedida. Como observado por [Wilson et al., 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 grande interesse neste domínio se deve à hipótese de que notícias de caráter positivo ou negativo, relacionadas com uma companhia pode afetar o desempenho financeiro da companhia em bolsas de valores. Assim, a polaridade de um documento de natureza financeira poderia ser usada para ajudar a prever tendências relacionadas com o desempenho de uma companhia. Com isso, identificar a polaridade de documentos financeiros é a tentativa de fazer previsões nos mercados de ações [Azar, 2009, Bollen et al., 2010, Devitt and Ahmad, a, Devitt and Ahmad, b] em que notícias positivas estão relacionadas a uma grande mudança no mercado e notícias negativas a pequenas mudanças, as quais são mais freqüentes.

O desafio torna-se maior, pois domínios de filmes e de produtos os autores dos textos avaliam com nota, diferente dos textos financeiros.[Azar, 2009].

Trabalhos sobre polaridade em domínio financeiro
[Azar, 2009, Bollen et al., 2010, Schumaker and Chen,] utilizam a premissa que o documento é a respeito de uma única entidade, porém isso não é verdade uma vez que diversos documentos citam duas ou mais entidades.

Nota-se na tabela 1, duas entidades em um documento: *Amazon* e *Apple*. A abordagem sugerida pelos trabalhos [Azar, 2009, Devitt and Ahmad, b],

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

infere a polaridade apenas para uma entidade já pré-determinada. Entretanto, o nível de polaridade para cada entidade é diferente. O texto referente a Amazon é positivo, pois afirma que o seu produto superou em vendas o da Apple. Para a Apple, o texto é negativo uma vez que a venda de seu dispositivo caiu e gerou um prejuízo financeiro.

Neste trabalho de mestrado, o foco é de identificar as entidades presentes no documento textual, verificar quais fragmentos de textos são relacionados a cada entidade e detectar a polaridade em relação a cada entidade.

O restante da proposta está dividido da seguinte maneira. Na seção 2 os objetivos são apresentados. Na seção 3 os trabalhos relacionados. Na seção 4 descreve a metodologia empregada no trabalho. Na seção 5 apresentaalguns resultados preliminares alcançados nesse trabalho. Na seção 6 mostra o cronograma de atividades.

2 Objetivos

2.1 Geral

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

particular.

2.2 Específicos

O trabalho possui os seguintes objetivos:

1. Criar coleção de documentos com múltiplas entidades;
2. Fazer pesquisa bibliográfica envolvendo análise de sentimentos em particular no domínio financeiro;
3. Sugerir métodos para detectar entidades no texto;
4. Determinar trechos dos textos associados a cada entidade;
5. Determinar a polaridades desses trechos

3 Revisão Bibliográfica

A análise de sentimentos é utilizado em diversos domínios de textos como resenhas de filmes e de produtos. Em [Pang et al., 2002], o objetivo do trabalho foi demonstrar como a classificação de documentos correspondente a sua polaridade é um trabalho tão simples quanto classificá-los com base em seus tópicos. Os experimentos foram realizados com base nas resenhas de filmes do site *Internet Movie Datavase* (IMdB), em que apenas os documentos que continham uma nota associada ao texto do documento foram utilizados. Para implementação da solução, três técnicas foram utilizadas: Naive Bayes, Máxima Entrópia e Support Vector Machine (SVM). Os resultados apresentados demonstram que a classificação dos documentos por sua polaridade é tão desafiante quanto por tópicos, isso porque as características observadas nos textos apontam ironias ou frases em que palavras com cunho positivo são usadas para uma frase negativa, definido como contextualização dos termos. Além disso, notou-se que o uso dessas técnicas é comparável com a classificação feita por pessoas e o método Naive Bayes apresentou uma performance pior do que os demais.

Baseado nesse problema de contextualização de termos, [Wilson et al., 2005] usa uma abordagem em que explora característica de frases para analisar o sentimento dos textos. A base de dados usada foi *Multi-perspective Question Answering* (MPQA), o seu conteúdo são documentos em língua inglesa da imprensa global. Outro fator importante, os documentos são detalhados e o conteúdo com 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. Um exemplo:

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

O termo Trust expressa um sentimento positivo, porém no contexto do texto, a palavra não é usada para expressar um sentimento e sim, o título da entidade. Por fim, o trabalho apresenta que termos classificados como positivos e negativos são comuns em frases neutras. Diferente dos dois trabalhos citados, [Yi et al.,] afirma que a polaridade de um documento deve ser com base em diferentes tópicos dentro do mesmo texto.

[Azar, 2009] faz uma análise entre a polaridade e as reações do mercado de ações, para verificar se é possível aplicar o algoritmo de polaridade no domínio financeiro para outro domínio. O autor usou como base a *Reuters Key Developments Corpus*, a qual contém notícias entre o período de 1998 e 2009. Porém, o autor utilizou somente companhias com mais de 20 notícias. Seus resultados foram obtidos com o uso de técnicas de processamento de linguagens naturais, além de Árvores de Decisão e SVM. O autor concluí que é possível aprender sobre a linguagem dos textos financeiros, termos mais usados e mais impactantes para mensurar a polaridade. Além disso, as técnicas utilizadas são tão boas quanto a avaliação de pessoas. Mas a análise de sentimento em domínio financeiro não pode ser usado em outros domínios.

Em [Bollen et al., 2010], o *Twitter* foi usado como base para experimentação. Com o uso das ferramentas *OpnionFinder*, mensura positivo e negativo, e *Google-Profile of Mood States* (GPOMS), o autor tem como objetivo antecipar o mercado de ações financeiras. Os resultados mostram que mudanças no estado emocional do público em análise de um grande volume de dados do *Twitter* mostram impactos dias depois no mercado financeiro. Seus resultados foram alcançados com simples técnicas de processamento de linguagem.

Os trabalhos [Devitt and Ahmad, a, Devitt and Ahmad, b] referência o trabalho de Darwin que categorizou em um conjunto finito as emoções básicas do homem: raiva, medo, tristeza, entre outros. Outra referência é a delimitação dos 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 consiste em construir um grafo que representa o texto todo, em que os nós são termos do texto com seus respectivos valores de polaridade de acordo

com a ferramenta SentiWordNet (ferramenta que mensura os termos em positivos e negativos de acordo com a WordNet). Com isso, sua abordagem para classificar a polaridade baseia-se em um conjunto léxico com termos referentes a positivo e negativo, junto com o apoio da teoria de Darwin tenta identificar quão intenso é esse sentimento.

Em [Schumaker and Chen,], o autor analisa notícias financeiras com base em diferentes reprsentações textuais: *bag of words*, sintagmas nominais e nome de entidades. Mesmo sem verificar a polaridade do texto, a importância do trabalho consiste em que o autor demonstra que modelos contendo termos de artigos e preços das ações no momento em que o artigo foi publicado tem uma grande acerto na estimativa dos preços da ações.

Os trabalhos citados dentro do domínio financeiro têm como característica comum, a análise de sentimentos em apenas uma entidade pré-determina ou tendo como consideração apenas uma entidade contida no texto.

4 Metodologia

O trabalho será implementado em três etapas: a primeira etapa é a criação de uma coleção para os experimentos. Uma fonte de dados é a Reuters Key Developments Corpus, essa base contém notícias sobre mercados de ações. Porém, é necessário solicitar o acesso para o seu uso. Uma possível fonte de dados é a Multi-perspective Question Answering (MPQA), a base é rica em termos com forte carga de emoção. Porém, é necessário filtrar uma vez que as 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.

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, a execução de algoritmos para a extração de terminologias, com isso poderemos verificar os termos relavantes no domínio financeiro e usar no cálculo da polaridade. Os algoritmos a serem utilizados neste trabalho são técnicas de Naive Bayes, Árvores de decisão e SVM. 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 é a *Beautiful Anaphora Resolution Toolkit* (BART). A BART utiliza-se de referência anafóras para verificar no documento em que trechos tal entidade é citada. Nós usaremos técnicas de Nayve Bayes para identificar tais fragmentos. Após a extração dos fragmentos de texto que uma entidade aparece no texto, 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. O uso de máquinas de apredizagem como SVM, métodos de classificação como Árvores de Decisão e modelos Naive Bayes serão utilizados.

A última etapa do nosso trabalho é a avaliação nos resultados. O uso de uma coleção com documentos pré-classificados em positivos, negativos e neutros 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

A fim de validar a proposta, 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*, tal lista contém as 2000 maiores companhias globais, caso o termo esteja, contabiliza-se uma entidade no documento.

A tabela 2 mostra o total de documentos com o número de entidades citadas no texto. Nota-se que o número maior de documentos apresentam cinco entidades em seus textos, porém 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.

A tabela 3 descreve as ações financeiras de várias entidades: *Google*, *IBM*, *Intuitive Surgical Inc.*, *JDS Uniphase Corp.*, *Marathon Petroleum* e *Microsoft*. O texto demonstra a queda nas ações do Google e alta nas ações da Microsoft, além de descrever outras entidades de ramos distintos. Entretanto, é possível inferir que neste documento a polaridade para o Google é negativo e para a Microsoft é positivo. Portanto, o número de entidades em cada

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	12
16	5
17	3
18	4
19	5
21	1
22	1
27	1

Tabela 2: Documentos com o número de entidades citadas em seu texto

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 3: Exemplo de documento com descrição das ações no mercado financeiro

documento mostra que inferir polaridade para um texto a respeito de apenas uma entidade não é a melhor abordagem.

O segundo experimento consiste em verificar a polaridade para um conjunto de entidades em diversos documentos. Para isso, cinco entidades foram utilizadas: Apple, Microsoft, Google, Nokia e Samsung. Tais entidades representam empresas de um mesmo ramo de negócio: celulares. Em seguida, selecionou-se um conjunto de cem páginas em que pelo menos três entidades são citadas. A polaridade foi feita de forma manual, pois essa análise poderá ser usada como treino em resultados futuros.

		Microsoft		Google		Nokia		Samsung	
		POS	NEG	POS	NEG	POS	NEG	POS	NEG
Apple	POS	2	7	3	0	6	9	4	2
	NEG	1	1	4	2	1	0	4	1

Tabela 4: 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	6	5	3	6	7	4	3	2
	NEG	5	0	3	4	0	1	1	3

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

A tabela 4 relaciona a entidade Apple com as demais entidades e a tabela 5 refere-se a Microsoft com as demais entidades. Nota-se a divisão em quatro 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, isso significa que existe 7 documentos entre essas 100 páginas em 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 Nokia, isso porque tais entidades são as maiores concorrentes entre seus setores.

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.

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

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

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

[Pang et al., 2002] Pang, B., Lee, L., and Vaithyanathan, S. (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 and Chen,] Schumaker, R. P. and Chen, H. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS).*

[Wilson et al., 2005] Wilson, T., Wiebe, J., and Hoffmann, P. (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.*

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