Applying Textmining to Classify News About Supply and Demand in the Coffee Market

P. O. Lima Júnior, L. G. Castro Júnior and A. L. Zambalde

***Abstract*— This work verifies the feasibility of text classification using supervised machine learning method to promote the web news monitoring on factors that impact supply and demand for the coffee market. To this end, a device was develop that enables the empirical evaluation of the Naive Bayes method to sort news collected from the web according to the categories: positive or negative to supply and to demand. The tests show the feasibility of Naive Bayes classifier to identify factors that affect supply and demand in coffee market.**

***Keywords*— Textmining, Machine Learning, Coffee Market.**

I. INTRODUÇÃO

APESAR de ser um setor expressivo para a composição do Produto Interno Bruto (PIB) do Brasil, a cafeicultura é uma atividade de risco para as empresas do setor, pois é sensível as variações de oferta e demanda mundial de café. Vários fatores influenciam oferta e demanda o que afeta o preço do café nos mercados e causa impactos em diferentes setores de sua cadeia produtiva.

A oscilação de preço do café no mercado atinge diretamente os agentes e decisões sobre comercialização e gerenciamento de risco, o que exige atenção às mudanças nos fundamentos da oferta e demanda [1]. Portanto, o monitoramento de fatores que impactam oferta e demanda é uma tarefa estratégica para organizações do setor.

Assim, a medida que aumenta o número de notícias sobre eventos que permitem identificar tendências, estimativas de produção, consumo e outros fundamentos do mercado de café, a possibilidade de extrair informação de textos torna-se relevante para a cafeicultura. O trecho de notícia extraída da *web*: "Baixa produção de café arábica na Colômbia pode favorecer exportação brasileira" é uma evidência qualitativa que contribui para análise de diminuição da oferta mesmo que suas implicações financeiras não sejam conhecidas de imediato.

Entretanto, o trecho citado como exemplo, assim como outros que influenciam o mercado de café, está em meio a milhões publicados na web diariamente. Lidar com este volume de informações não é uma tarefa trivial, consome recursos humanos, tempo e restringe a análise à capacidade de busca e leitura de especialistas.

P. O. Lima Júnior, CEFET-MG, Nepomuceno, MG, plima@nepomuceno,cefetmg.br
L. G. Castro Júnior, Universidade Federal de Lavras, Lavras, MG, lgcastro@dae.ufla.br
A. L. Zambalde, Universidade Federal de Lavras, Lavras, MG, zamba@dcc.ufla.br

A automatização do processo, esbarra em obstáculos no campo sintático ou ruídos nos dados, e semântico – ambiguidade da linguagem e ausência de contexto. Estes obstáculos suscitam técnicas de Mineração de Dados (*Datamining*) para extração de informação a partir de grande volume de dados.

Entre as técnicas da Mineração de Dados estão: classificação, previsão numérica, associação e agrupamento (*clustering*). Segundo [2], uma tendência dos avanços em Mineração de Dados é a Mineração Textual, que aplica as mesmas técnicas para informação textual resultando em funções para análise de texto. A técnica de classificação, neste contexto, é a tarefa de separar textos em classes pré-definidas. Uma das abordagens para esta tarefa é o uso de algoritmos de aprendizado de máquina, que a partir de treinamento com um conjunto de textos previamente classificados criam modelos matemáticos que permitem classificar novos textos.

Dentre os algoritmos estudados e utilizados como classificadores, destacam-se os pertencentes a três classes [3] [4] [5]: *Naïve Bayes*, Árvore de Decisões e *Support Vector Machines*.

Com estas considerações, este trabalho tem como objetivo verificar a contribuição da classificação textual, por aprendizado de máquina, para monitorar notícias na web que exercem influência no mercado de café. Para isso foi desenvolvido um protótipo, baseado em Mineração Textual, para coleta de notícias da web e classificação automática quanto ao impacto para oferta e demanda.

A pesquisa foi realizada, seguindo a metodologia *Design Science Research*, no Bureau de Inteligência Competitiva do Café (BIC), projeto do Centro de Inteligência em Mercados da Universidade Federal de Lavras que tem entre seus objetivos, produzir, a partir de notícias sobre o mercado de café, relatórios com informações e análises sobre o setor cafeeiro.

A banco de dados existente no BIC com notícias coletadas da web de 2011 a 2015 configura um ponto inicial como base de treinamento para algoritmos classificadores e viabiliza a abordagem por aprendizado de máquina.

O experimento apresenta o desempenho do método *Naive Bayes* em três situações para classificação de notícias. Os resultados mostram que é viável o uso de classificador pelo BIC para automatização da coleta e classificação, ganho de escala e redução de recursos humanos.

O artigo está organizado da seguinte forma: como referencial teórico a Seção 2 apresenta Mineração Textual e a Seção 3 Oferta e Demanda no mercado de café, também como referência para a criação das categorias para classificação de notícias. A Seção 4 descreve a metodologia utilizada. Na Seção 5 são apresentadas as etapas da pesquisa de acordo com a metodologia adotada. A avaliação dos resultados é apresentada na seção 6 e a conclusão na Seção 7.

II. MINERAÇÃO TEXTUAL

Mineração Textual é uma área interdisciplinar com fundamentos teóricos da Estatística, Ciência da Computação e Inteligência Artificial. Explora a descoberta e extração de informação e conhecimento válido a partir de grande volume de dados textuais pela aplicação de métodos de Processamento de Linguagem Natural, Aprendizado de Máquina e Recuperação da Informação. Promove a eficiência das organizações para coleta, processamento e análise de informação [2] [6].

A área foi impulsionada principalmente pela crescente produção de textos na *web*. No campo digital faz uso de técnicas para lidar com dados não estruturados ou semi-estruturados – dados não organizados em um modelo conceitual que facilita sua recuperação, como em um banco de dados. Estas técnicas incluem algoritmos para classificação, sumarização, identificação de entidades e *text clustering*.

O estudo proposto neste trabalho refere-se a classificação de texto para organização de documentos. Esta tarefa tem como objetivo classificar textos de acordo com categorias pré-definidas [7]. Quando existe um conjunto de dados previamente rotulados com as categorias a tarefa é considerada supervisionada – os dados são utilizados para treinar algoritmos de aprendizado de máquina.

A classificação supervisionada é dividida em duas fases distintas: treinamento e classificação. Na fase de treinamento, um algoritmo com a técnica específica é treinado com um conjunto de textos previamente classificados. Na fase de classificação o algoritmo classifica novos textos de acordo com as categorias aprendidas na fase anterior.

Estudos mostram que técnicas existentes alcançam eficiência e precisão satisfatórias quando treinados em grades conjuntos de dados [8], [3], [9]. O problema, conforme [10], é que em aplicações reais, os recursos humanos necessários para a produção manual da base de treinamento dificultam ou inviabilizam o processo.

O foco deste trabalho é o método Naive Bayes selecionado principalmente pela sua simplicidade e eficácia [11] e disponibilidade de ferramentas. Os métodos Bayesianos (*Naïve Bayes*) constroem um modelo probabilístico baseado na ocorrência de palavras nas diferentes categorias. O algoritmo classifica o documento baseado na probabilidade deste pertencer a determinada categoria conforme as palavras presentes no texto [12].

A definição das categorias para fase de aprendizado, realizada pela equipe do BIC de acordo com a cadeia produtiva do café, é descrita na próxima seção.

III. OFERTA E DEMANDA E GERENCIAMENTO DE RISCO NO MERCADO DE CAFÉ

Para produtores, exportadores e consumidores, a variação de preço no mercado de café afeta diretamente decisões de comercialização associadas a gerenciamento de risco [13] [14] [15]. Um dos mecanismos utilizados pelos agentes para o gerenciamento de risco e proteção contra eventuais perdas diante da volatilidade do preço, é o mercado de derivativos, via operações de *hedge* [16], na qual troca-se o risco de variação do preço ao comercializar o produto exclusivamente no mercado físico pelo risco de variação da diferença entre o preço físico e o preço futuro. Essa diferença entre o preço da *commodity* no mercado físico, na praça local de comercialização e o preço futuro para determinado mês de vencimento do contrato no mercado futuro é denominado base [16].

Entretanto, o *hedge* não é uma operação trivial, há incerteza quanto ao momento adequado para abertura e encerramento da operação e existe a volatilidade provocada por vários fatores, que vão desde boatos ou expectativas sobre o clima sem respaldo técnico científico [17] até a ação de especuladores. Este cenário, exige atenção às mudanças nos fundamentos da oferta e demanda [1].

Estudos sobre volatilidade [18], [14], [15], [17], [19], [20] e efetividade do *hedge* [21], [22] apontam que períodos específicos de alta volatilidade são explicados por eventos relacionados ao mercado, tais como: quebra de safra, condições climáticas e adversidades para comercialização do produto.

Conforme [17], [23] e [24] fatos relevantes sobre o mercado de café são divulgados na mídia e a compreensão de eventos que provocam volatilidade no preço do café e afetam a dinâmica da sua cadeia produtiva é relevante para a tomada de decisões. Portanto, o monitoramento de fatores que impactam oferta e demanda é uma tarefa estratégica para organizações do setor.

A análise de eventos é uma das tarefas para identificar tendências, estimativas de produção, consumo e outros fundamentos do mercado de café [23]. Na cafeicultura, [18] apontam que notícias negativas sobre o mercado de café afetam diretamente a produção e contribuem, de forma expressiva, na volatilidade da base do produto. Além disso, surtos de alta volatilidade têm duração limitada [17] e choques na base, positivos ou negativos, demoram um tempo considerável para se dissiparem [18].

A partir da modelagem estatística da série de retorno diário do café futuro na Bolsa de Mercadorias e Futuros do Brasil (BM&F), [24] observa assimetria a boas e más notícias e ao analisar 15 picos de volatilidade do modelo da série, aponta relevância para fatores relacionados à produção: ciclo bianual e eventos climáticos, reforçando que a volatilidade no mercado futuro é vulnerável a esses fatores.

[17] ressaltam que a incerteza gerada pela volatilidade ao invés de incentivar a utilização de *hedge* para fixação de preço em muitos casos afasta os agentes em função da dificuldade de bancar as diferenças de margens, que exigem um aporte significativo de capital de curto prazo.

Assim, identificar e monitorar fatores que afetam oferta e demanda de café no mundo e entender sua relação com preço e volatilidade representa vantagem competitiva para empresas do setor, à medida que auxiliar sobre o melhor momento de ajustar posições no mercado futuro para evitar flutuações de receitas, sabendo que efeitos na base duram um período significativo.

IV. *DESIGN SCIENCE RESEARCH*

Por buscar entendimento sobre fenômenos criados pelo homem (publicação de dados na *web*) com o uso de tecnologia (Mineração Textual), para resolução de problema em um contexto específico (cafeicultura), esta pesquisa segue o

paradigma *Design Science* pelo método *Design Science Research*.

Design Science Research tem como objetivo estudar, pesquisar e investigar o artificial e seu comportamento sob a perspectiva acadêmica e organizacional [25]. A ideia central do método é que a aquisição de conhecimento e a solução de um problema acontecem pela construção e aplicação de um artefato: constructos, modelos, métodos, instanciações e aprimoramento de teorias existentes.

Para [26], *Design Science Research* é a análise do uso e desempenho de artefatos projetados para compreender, explicar e melhorar o comportamento de determinados aspectos na área de sistemas de informação.

O modelo adotado neste trabalho é adaptado do proposto por [26], com contribuições de [27], por ter suas raízes em Sistemas de Informação. Inclui 5 etapas: Entendimento do problema, Sugestões, Desenvolvimento, Avaliação e Conclusão.

A investigação tem início pelo conhecimento de um problema ou oportunidade de pesquisa na primeira etapa. Deve-se identificar e compreender o problema a ser solucionado e definir qual desempenho necessário para o sistema em estudo. A saída desta etapa é uma proposta formal ou não, com evidencias da situação problemática, caracterização do ambiente externo e interação com o artefato, definição de métricas e critérios de validação do artefato [27].

Na etapa de sugestão são elaborados um ou mais modelos de tentativa para a solução do problema a partir da existência de conhecimento/teoria de base. Predomina o método abdutivo que exige processo criativo e conhecimento prévio para proposta de soluções [26]. A saída desta etapa são tentativas e suas comparações com o intuito de resolver o problema.

Os artefatos propostos são construídos na etapa de desenvolvimento, e avaliados na quarta etapa. Caso o artefato seja inadequado durante as etapas de desenvolvimento ou avaliação, retorna-se à primeira etapa para revisão do entendimento do problema. Este ciclo denominado circunscrição é fundamental para a construção incremental de conhecimento.

Na etapa de conclusão resultados são consolidados e registrados e podem ser retroalimentados no processo já que pode-se concluir incompletude ou insuficiência da conscientização do problema comprometendo o desempenho do artefato. Neste processo é possível identificar lacunas na teoria e recomeçar o processo.

As etapas do método conduzidas nesta pesquisa foram realizadas no Centro de Inteligência em Mercados com profissionais do Bureau de Inteligência do Café (BIC).

V. ETAPAS DO MÉTODO

O Bureau de Inteligência Competitiva do Café (BIC) é um projeto do Centro de Inteligência em Mercados da Universidade Federal de Lavras que utiliza dados da *web* para auxiliar a produção de relatórios com delineamento de cenários para tomada de decisões no mercado de café. Uma equipe, formada por especialistas, monitora notícias sobre o mercado diariamente, classifica as relevantes para a cadeia produtiva do café e armazena em um banco de dados para produção de relatórios. Neste ambiente, juntamente com os especialistas, foi possível verificar o que é extraído da *web*, como o resultado deve ser classificado e o que é analisado para oferta e demanda.

Entendimento do Problema

O banco de dados utilizado neste trabalho é composto por 3117 notícias em inglês, coletadas manualmente da web por especialistas do Bureau no período de 01/01/2011 a 31/12/2015. São armazenados título, fonte, conteúdo e categoria que indica o setor da cadeia produtiva para o qual a notícia é relevante.

Os relatórios são produzidos mensalmente, porém todo o processo é restrito a capacidade dos especialistas para coletar de diversas fontes, ler e classificar as notícias relevantes, o que consome tempo e recursos humanos. Nem sempre o número de notícias é suficientes o que pode comprometer a abrangência do relatório. As notícias não são classificadas em categorias específicas para monitoramento de oferta e demanda.

Neste cenário, temos o problema enunciado como pergunta: É possível auxiliar o monitoramento de oferta e demanda no mercado de café por meio de Mineração Textual de notícias da web?

Sugestão

O projeto para construção do protótipo (artefato) segue modelos tradicionais para extração de conhecimento [28], [29], *Business Intelligence* [30], [31] e [32] e Sistemas baseados em Mineração Textual para Inteligência Competitiva [33], [34] e [35].

O protótipo foi dividido por tarefas em 4 módulos: supervisão, coleta, pré-processamento e classificação. O banco de dados de notícias do BIC formou a base para treinamento de algoritmo para classificar novas notícias extraídas pelo módulo de coleta. Antes de iniciar processo, as notícias existentes no banco de dados são classificadas manualmente através do módulo de supervisão.

Para identificar e monitorar notícias publicadas na web com evidencias qualitativas que impactam oferta e demanda de café, as classes para classificação textual foram definidas juntamente com os especialistas do BIC. As classes são apresentadas na Tabela I, bem como a descrição e um exemplo de trecho que caracteriza cada classe.

TABELA I
SEGMENTOS DA CADEIA PRODUTIVA DO CAFÉ

Classe	Descrição	Exemplo
Oferta +	Evidências qualitativas que indicam aumento de oferta	*"Colombia coffee production recovers"*
Oferta -	Evidências qualitativas que indicam diminuição de oferta	*"Drought might harm Uganda's coffee harvest"*
Demanda +	Evidências qualitativas que indicam aumento de demanda	*"Lavazza announces manufacturing facility in India"*
Demanda -	Evidências qualitativas que indicam diminuição de demanda	*"Price of coffee may give bigger jolt than caffeine"*

O módulo de coleta é usado para recuperar notícias da web por consulta de termos específicos em motores de busca. A vantagem é ter uma ferramenta de busca amplamente utilizada

e aprimorada especificamente para este fim por empresas do mercado.

A saída deste módulo são notícias brutas, com ruídos em nível sintático: elementos de linguagem de marcação, códigos e elementos de estilo, e semântico: texto irrelevante à notícia, como texto publicitário ou de navegação no site, e irrelevância da própria notícia para o contexto do sistema – Cafeicultura. Estes problemas são abordados no módulo de pré-processamento.

O módulo de pré-processamento tem como objetivo preparar o conteúdo apenas com o essencial, assim inclui tarefas como retirada de *tags HTML*, *scripts* e outros ruídos presentes em texto extraído da *web* e não relevantes para a classificação. Outra tarefa é a remoção de *stopwords* – palavras frequentes que não são significativas para representar uma classe como artigos, pronomes, preposições e advérbios. E para reduzir a variação dos termos, por exemplo: *producer*, *producers*, *produce*, *pruducing*, para uma única representação aplica-se o procedimento denominado *stemming* [36].

Posteriormente, os textos com classes conhecidas (rotulados) são convertidos em uma representação numérica constituindo um modelo para, no módulo de classificação, predizer a classe de novos textos coletados da web. Todas as notícias passam por este processo como um treinamento para o módulo de classificação.

Desenvolvimento

O sistema foi desenvolvido como uma aplicação Java. O processo foi realizado com a utilização da interface de programação da ferramenta WEKA [37] que contém algoritmos para pré-processamento e classificação de dados.

Para o módulo de coleta foi utilizado o motor de busca *Google Search*. Foi criado um dicionário de termos comuns ao contexto da cafeicultura, definidos pelos especialistas e pelo estudo da base de dados existente incluindo a combinação de palavras relativas a tópicos como: produção, indústria, empresas, países, clima, doenças e pragas.

Assim, para automatizar a tarefa de coleta de dados, o sistema percorre um arquivo com a lista de termos e, através da interface do motor de busca, recupera notícias resultantes da consulta para cada termo.

No módulo de pré-processamento, foi utilizada a biblioteca de classes em Java, Apache Tika [38], que contém funções específicas para limpeza de texto. Foi considerada a lista de *stopwords*, em Inglês, disponível em http://www.ranks.nl/stopwords. O algoritmo para a tarefa de *stemming* foi o Snowball [39], utilizado para língua inglesa.

O filtro utilizado para tratamento de texto da ferramenta WEKA é o *StringToWordVector*. Sua função é converter texto em um conjunto de atributos representando a ocorrência de cada palavra no texto. Este conjunto é armazenado em um vetor posteriormente utilizado na fase de treinamento.

Para determinar a relevância de palavras foi utilizado o método *Term Frequency Inverse Document Frequency* (*TF-IDF)* disponível na ferramenta WEKA. Este método considera a frequência do termo no documento e sua relevância para todo o conjunto de documentos. Conforme [8]:

$$TF - IDF_{t,d} = TF_{t,d} \times IDF_t \quad (1)$$

Onde,

$$IDF_t = \log \frac{N}{DF_t}, \quad (2)$$

$TF_{t,d}$ é o número de ocorrências do termo t no documento d, N é número de documentos no conjunto e DF_t é número de documentos no conjunto que contém o termo t.

O modelo para classificação foi gerado a partir das notícias as classes atribuídas no módulo de supervisão.

Para o módulo de classificação, o banco de dados foi dividido em dois conjuntos de dados: $\Omega_t = *d_1, \ldots d_{|\Omega t|}*$, onde $|\Omega_t|$ representa o número de notícias compreendido entre duas datas para treinamento. E o conjunto: $\Omega_c = *d_1, \ldots d_{|\Omega c|}*$, onde, $|\ \Omega_c\ |$ representa o número de notícias compreendido entre duas datas para classificação, usado para avaliar o desempenho dos classificadores. As categorias para classificação são representadas pelo conjunto: $C = *$*"Oferta+", "Oferta-", "Demanda+", "Demanda-"*.

A criação do conjunto de treinamento foi realizada pelos especialistas na etapa de supervisão, onde classificaram aleatoriamente 553 notícias coletadas de 01/01/2011 a 31/12/2015. A classificação resultou em notícias em Inglês rotuladas conforme Tabela II.

TABELA II
NOTÍCIAS CLASSIFICADAS PELOS ESPECIALISTAS PARA OFERTA E DEMANDA

Classes	Número de Notícias
Oferta+ (aumento de oferta)	83
Oferta- (diminuição de oferta)	73
Demanda+ (aumento de demanda)	149
Demanda- (diminuição de demanda)	48
Nenhuma	200

O desbalanceamento da base de treinamento é uma característica que interfere o desempenho da classificação e deve ser tratada por técnicas de balanceamento [40]. Portanto para o balanceamento da base de treinamento foi utilizado o método *Resample* da ferramenta WEKA em todos os testes.

O algoritmo classificador utilizado é a versão *Naive Bayes* [41] codificada na ferramenta WEKA com parâmetros padrão.

O experimento foi executado em uma aplicação programada em Java, desenvolvida para conectar ao banco de dados do BIC, pesquisar os conjuntos de notícias, treinar o algoritmo via interface de programação WEKA e utilizá-lo para classificar outros conjuntos de notícias.

VI. AVALIAÇÃO

Para avaliar o desempenho do classificador foi realizada uma comparação entre sua classificação e a classificação realizada pelos especialistas para um mesmo conjunto de notícias. A Tabela III apresenta os resultados possíveis na comparação entre classificador e especialista para atribuir uma classe C_i a uma notícia.

Na Tabela III, TP_i significa que o classificador convergiu com o especialista na decisão de classificar na classe C_i, enquanto FP_i é um falso positivo, classificador errou a classe atribuída pelo especialista.

TABELA III
COMPARAÇÃO ENTRE CLASSIFICADOR E ESPECIALISTA PARA UMA CATEGORIA C_i

Classe C_i		Especialista	
		Sim	Não
Classificador	Sim	TP_i (*True Positive*)	FP_i (*False Positive*)
	Não	FN_i (*False Negative*)	TN_i (*True Negative*)

Fonte: Adaptado de [3]

Conforme [3] a avaliação experimental em classificação textual usualmente mede eficácia – capacidade de tomar a decisão correta de classificação, principalmente por sua característica subjetiva. Assim, foram consideradas as medidas de Precisão (π) e Revocação (ρ) propostas pelo autor:

$$\pi i = \frac{TPi}{TPi+FPi} \quad (3)$$

$$\rho i = \frac{TPi}{TPi+FNi}. \quad (4)$$

As fórmulas denotam respectivamente as probabilidades de estar correta a decisão de classificar uma notícia aleatória N_x em uma classe C_i e de ser tomada a decisão correta de classificar em uma classe C_i uma notícia N_x que deve ser classificado nesta classe.

E a Medida F – a média harmônica entre Precisão e Revocação:

$$F = \frac{2*\pi i*\rho i}{\pi i+\rho i} \quad (5)$$

Desta forma é possível observar o comportamento do classificador para classes com diferentes quantidades de notícias para treinamento. Para a avaliação, foram realizados experimentos em três etapas.

O primeiro teste teve como objetivo analisar o desempenho do classificador *Naive Bayes* com o conjunto de teste formado pelas 553 notícias classificadas pelos especialistas, conforme Tabela II. Para avaliação, foi utilizada a técnica de Validação Cruzada, disponível na ferramenta WEKA, que separa nove partes do conjunto de notícias para treinamento e uma parte do conjunto para teste e avaliação do modelo gerado pelo algoritmo. O resultado é apresentado na Tabela IV.

TABELA IV
VALIDAÇÃO CRUZADA PARA CLASSIFICAÇÃO DE OFERTA E DEMANDA

Categoria	Taxa TP	Taxa FP	Precisão	Revocação	Medida F
Oferta+	0,653	0,061	0,628	0,653	0,641
Oferta-	0,893	0,127	0,754	0,893	0,817
Demanda+	0,531	0,07	0,5	0,531	0,515
Demanda-	0,538	0,016	0,778	0,538	0,636
Nenhuma	0,608	0,15	0,686	0,608	0,645

Legenda: TP Taxa – Taxa de Acerto de *True Positive* (Verdadeiro Positivo), FP Taxa – Taxa de Acerto de Falso Positivo.

Do total de 553 notícias, 379 foram classificadas corretamente (68,53%). O desempenho geral é aceitável entretanto o resultado mostra que as classes Oferta- e Demanda+ possuem melhor precisão enquanto as outras classes apresentam desempenho que compromete o resultado.

Em outro experimento, a base de treinamento foi composta por 492 notícias em inglês coletadas entre 01/01/2011 e 31/12/2013 e o conjunto de teste formado por 91 notícias coletadas no período de 01/01/2014 a 31/12/2014, também em inglês, distribuídas conforme Tabela V. As notícias do ano de 2014 foram classificadas pelo método *Naive Bayes* e comparadas com a classificação do especialista.

TABELA V
DISTRIBUIÇÃO DE NOTÍCIAS PARA TREINO E TESTE

Categorias	Treino	Teste
Oferta+	74	9
Oferta-	65	8
Demanda+	112	28
Demanda-	41	6
Nenhuma	200	24

O resultado distribuído por classes é apresentado na Tabela VI.

TABELA VI
RESULTADO DA CLASSIFICAÇÃO DISTRIBUÍDO POR CATEGORIAS

Categoria	TP	FP	FN	Precisão	Revocação	Medida F
Oferta+	2	3	7	0,400	0,222	0,286
Demanda+	21	10	7	0,677	0,750	0,712
Oferta-	4	4	4	0,500	0,500	0,500
Demanda-	4	9	2	0,308	0,667	0,421
Nenhuma	8	10	16	0,444	0,333	0,381

Legenda: TP – *True Positive* (Verdadeiro Positivo), FP – Falso Positivo, FN – Falso Negativo.

Do total de 75 notícias, 39 foram classificadas corretamente (52,00%). O resultado é prejudicado pelas classes Demanda-, Oferta+ e Nenhuma que apresenta baixo nível de Revocação.

Na última etapa, o objetivo foi classificar notícias coletadas da web pelo protótipo e comparar com a classificação realizada pelo especialista. Para isso a base de teste foi criada pela consulta automática, no módulo de coleta, retroativa mês a mês, de Janeiro de 2011 a Dezembro de 2015, por 132 termos definidos pelos especialistas, o que retornou 35.083 notícias válidas. A base de treinamento foi composta por 353 notícias de 01/01/2011 até 31/01/2015, conforme Tabela II.

Diante da inviabilidade de realizar a classificação manual de 35.083 notícias, pelos especialistas, para avaliar o desempenho do classificador, foram selecionadas as notícias com data de publicação entre 01/01/2014 e 31/03/2014, período onde aconteceu um choque de volatilidade significativo aumentando o preço do café, conforme mostra Fig. 1.

Figura 1. Variação do Preço do Café na Bolsa de NY (ICE).

E, buscando compatilidade com fundamentos do mercado que ocasionam alta de preços, foram selecionadas noticias nesse período classificadas como Oferta – e Demanda + que sugerem respectivamente diminuição de oferta e aumento da demanda.

A Tabela VII apresenta algumas notícias recuperadas da web pelo módulo de coleta no período avaliado e classificadas pelo módulo classificador quanto a oferta e demanda.

TABELA VII
AMOSTRA DE NOTÍCIAS PARA DEMANDA POSITIVA E OFERTA NEGATIVA

Título	Classe
"Nespresso launches large-cup coffee machine for North America"	Demanda+
"GMCR Launches Lavazza Coffee In Keurig K-Cups For Home, Office"	Demanda+
"Nestle Launches Super Size Coffee Maker to Take on Keurig"	Demanda+
"Kenya expects 2013/14 coffee export earnings to fall on global prices"	Oferta-
"El Nino threatens to return, hit global food production"	Oferta-
"Drought Could Drain More Than Brazil's Coffee Crop"	Oferta-

A amostra classificada foi submetida à avaliação dos especialistas que apontou 25 notícias corretamente classificadas como Demanda+ do total de 45 do período (55,55%) e 35 corretamente classificadas como Oferta– em 61 (57,37%)

VII. CONCLUSÕES

A avaliação de técnicas de Mineração Textual normalmente é feita por comparação com o desempenho em bases de dados existentes. [42] apresenta o desempenho do método *Naive Bayes* em 45 bases de documentos diferentes, com a porcentagem média de 67,29 classificações corretas.

Tomando este resultado como parâmetro, e considerando as limitações da base de treinamento do BIC: desbalanceamento, quantidade de amostras e ruídos no texto, os testes mostram que o classificador *Naive Bayes* apresenta precisão, revocação e medida F relevantes para as categorias Oferta + no teste de validação cruzada e Demanda + no teste em período específico (2014), assim como para classificar notícias extraídas da web pelo módulo de coleta no período avaliado – 55,55% para Demanda + e 57,37% para Oferta -.

Porém a precisão deve ser aumentada pelo Módulo de Supervisão com ajustes realizados pelos especialistas após a coleta e classificação automática do conjunto de notícias extraídas da *web*, aperfeiçoando a base de treinamento com mais amostras para as classes com menor precisão, visando alcançar índices plausíveis para o Naive Bayes conforme [42] (acima de 80%). É necessário também criar uma base de notícias irrelevantes, o que pode ser realizado pelos especialistas no módulo de supervisão e aprimorar o módulo de pré-processamento apra eliminar ruídos e texto irrelevante.

Não é um resultado definitivo ao passo que o comportamento do classificador se altera com características do conjunto de treinamento e pré-processamento, assim os resultados mostram que apesar do baixo desempenho para algumas categorias específicas, é viável o uso de classificadores pelo BIC para automatização da coleta e classificação, ganho de escala e redução de recursos humanos, pois apresentou resultado promissor diante da possibilidade de diminuir as limitações que comprometem seu desempenho geral pelo incremento da base de treinamento no módulo de supervisão. Além disso há outros classificadores com desempenho superior demonstrados em [42] que devem ser testados com os dados do BIC.

Pela perspectiva tecnológica, o trabalho mostra a utilidade e eficácia do protótipo através de um método rigoroso para construção e validação. Entretanto, sua utilização efetiva no ambiente do BIC depende dos ajustes apontados nesta conclusão em trabalhos futuros, que também incluem: testes com outros classificadores, maior número de notícias para treinamento e base balanceada, inclusão de mecanismos para extrair entidades, categorias mais específicas visando um suporte mais robusto não só para monitoramento de notícias, mas também apoio a Inteligência Competitiva na Cafeicultura.

REFERÊNCIAS

[1] D. H. d. Oliveira *et al.*, "Panorama da cafeicultura de Coffea canephora: perspectivas para Brasil e Vietnã," presented at the VIII Simpósio de Pesquisa dos Cafés do Brasil, Salvador BA, 2013.

[2] R. Bose, "Competitive intelligence process and tools for intelligence analysis," *Industrial Management & Data Systems,* vol. 108, no. 4, pp. 510-528, 2008.

[3] F. Sebastiani, "Machine Learning in Automated Text Categorization," *ACM Computing Surveys,* vol. 34, no. 1, pp. 1-47, 2002.

[4] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999, pp. 42-49: ACM.

[5] H. Kang, S. J. Yoo, and D. Han, "Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews," *Expert Systems with Applications,* vol. 39, no. 5, pp. 6000-6010, // 2012.

[6] H. Chen, M. Chau, and D. Zeng, "CI Spider: a tool for competitive intelligence on the Web," *Decision Support Systems,* vol. 34, no. 1, pp. 1-17, 2002.

[7] T. Joachims, *Text categorization with support vector machines: Learning with many relevant features*. Springer, 1998.

[8] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge university press Cambridge, 2008.

[9] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, "Rcv1: A new benchmark collection for text categorization research," *The Journal of Machine Learning Research,* vol. 5, pp. 361-397, 2004.

[10] A. McCallum, K. Nigam, J. Rennie, and K. Seymore, "A machine learning approach to building domain-specific search engines," in *IJCAI*, 1999, vol. 99, pp. 662-667: Citeseer.

[11] M. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text classification using machine learning techniques," *WSEAS Transactions on Computers,* vol. 4, no. 8, pp. 966-974, 2005.

[12] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, 1998, vol. 752, pp. 41-48: Citeseer.
[13] Á. d. M. Barros and D. R. Aguiar, "Gestão do risco de preço de café arábica: uma análise por meio do comportamento da base," *Revista de Economia e Sociologia Rural,* vol. 43, no. 3, pp. 443-464, 2005.
[14] A. L. R. Mól, "Value-at-Risk da Base em Operações Vendidas de Hedge nos Contratos Futuros de Café Arábica na BM&F," *Interface,* vol. 5, no. 1, 2011.
[15] C. M. Santos, T. A. Farias, F. R. de Moura, and R. S. Mateus, "MERCADO FUTURO DE CAFÉ: UM ESTUDO DE CASO," *Registro Contábil,* vol. 3, no. 1, pp. 62-84, 2012.
[16] J. Hull, *Fundamentos dos mercados futuros e de opções*. Bolsa de Mercadorias & Futuros, 2005.
[17] R. Nunes, M. S. M. Saes, and J. A. Brando, "A volatilidade das cotações de café nas bolsas internacionais," presented at the XLII Congresso da SOBER, Cuiabá, 2004.
[18] E. S. Melo and L. B. Mattos, "Análise da volatilidade da base do café arábica para a mesorregião do sul de Minas Gerais DOI–10.5752/P. 1984-6606.2012 v12n29p95," *Revista Economia & Gestão,* vol. 12, no. 29, pp. 124-140, 2012.
[19] J. M. Fry, B. Lai, and M. Rhodes, "The interdependence of coffee spot and futures markets," *International Network for Economic Research Working Paper Series, Speyer, London,* 2011.
[20] Y. Zhang, "Forecasting of daily dynamic hedge ratio in agricultural and commodities' futures markets: evidence from Garch models," University of Southampton, 2012.
[21] R. E. Fontes, L. CASTRO JUNIOR, and A. F. Azevedo, "Base e risco de base da cafeicultura em Minas Gerais e São Paulo," *Artigos Técnicos de Derivativos Agropecuários. São Paulo, resenha BM&F,* no. 153, 2003.
[22] A. R. Pavão, "Análise do comportamento da base do Café Arábica: um estudo de caso do
município de Alpinópolis – MG," presented at the CONGRESSO BRASILEIRO DE ECONOMIA E
SOCIOLOGIA RURAL, Campo Grande, 2010.
[23] G. F. de Abreu, E. C. Silva, L. G. de Castro Junior, and P. H. A. Santos, "IDENTIFICAÇÃO DAS PRINCIPAIS TENDÊNCIAS PARA A PRODUÇÃO MUNDIAL DE CAFÉ," presented at the VIII Simpósio de Pesquisa dos Cafés do Brasil Salvador, BA, 2013.
[24] C. M. F. Martins, "A volatilidade nos preços futuro do café brasileiro e seus principais elementos causadores," 2015.
[25] N. Bayazit, "Investigating design: A review of forty years of design research," *Design issues,* vol. 20, no. 1, pp. 16-29, 2004.
[26] V. Vaishnavi and W. Kuechler, "Design research in information systems," 2004.
[27] N. Manson, "Is operations research really research?," *ORiON: The Journal of ORSSA,* vol. 22, no. 2, pp. 155-180, 2006.
[28] R. Moraes, J. F. Valiati, and W. P. Gaviao Neto, "Document-level sentiment classification: An empirical comparison between SVM and ANN," *Expert Systems with Applications,* vol. 40, no. 2, pp. 621-633, Feb 1 2013.
[29] M. Bramer, *Principles of data mining*. Springer, 2013.
[30] H. Baars and H.-G. Kemper, "Management support with structured and unstructured data—an integrated business intelligence framework," *Information Systems Management,* vol. 25, no. 2, pp. 132-148, 2008.
[31] R. Bose, "Advanced analytics: opportunities and challenges," *Industrial Management & Data Systems,* vol. 109, no. 2, pp. 155-172, 2009.
[32] J.-Y. Wu, "Computational intelligence-based intelligent business intelligence system: concept and framework," in *Computer and Network Technology (ICCNT), 2010 Second International Conference on*, 2010, pp. 334-338: IEEE.
[33] Y. Dai, T. Kakkonen, and E. Sutinen, "MinerVA: a decision support model that uses novel text mining technologies," in *Management and Service Science (MASS), 2010 International Conference on*, 2010, pp. 1-4: IEEE.
[34] Y. Dai, T. Kakkonen, and E. Sutinen, "MinEDec: a decision-support model that combines text-mining technologies with two competitive intelligence analysis methods," *International Journal of Computer Information Systems and Industrial Management Applications,* vol. 3, pp. 165-173, 2011.
[35] Y. Dai, T. Kakkonen, E. Arendarenko, D. Liao, and E. Sutinen, "MOETA: a novel text-mining model for collecting and analysing competitive intelligence," *International Journal of Advanced Media and Communication,* vol. 5, no. 1, pp. 19-39, 2013.
[36] S. M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, *Text mining: predictive methods for analyzing unstructured information*. Springer, 2010.
[37] G. Holmes, A. Donkin, and I. H. Witten, "Weka: A machine learning workbench," in *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*, 1994, pp. 357-361: IEEE.
[38] C. Mattmann and J. Zitting, *Tika in Action*. Manning Publications Co., 2011.
[39] M. Porter, "Snowball: A language for stemming algorithms," ed, 2001.
[40] F. d. R. N. Guimarães, "Descobrindo Padrões de Gêneros das Mensagens em Fóruns de Discussão de Ambientes Virtuais de Aprendizagem via Mineração de Texto," Dissertação de Mestrado, Universidade Federal de Lavras, 2015.
[41] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 1995, pp. 338-345: Morgan Kaufmann Publishers Inc.
[42] R. G. Rossi, R. M. Marcacini, and S. O. Rezende, "Benchmarking text collections for classification and clustering tasks," *Institute of Mathematics and Computer Sciences, University of Sao Paulo,* 2013.

Paulo de Oliveira Lima Júnior é graduado em Ciência da Computação pela Universidade Federal de Ouro Preto (UFOP), Ouro Preto, Minas Gerais, Brasil em 2000. Obteve o título de mestre em Computação Aplicada pelo Insituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, SP, Brasil em 2002. Possui Doutorado em Administração pela Universidade Federal de Lavras na área de Gestão de Negócios, Economia e Mercados. Atualmente é professor do Centro Federal de Educação Tecnológica de Minas Gerais no Departamento de Mecânica e Computação, Campus Nepomuceno, MG, Brasil. Suas principais áreas de pesquisas são: Mercados de derivativos, Mineração Textual e Análise de Sentimento.

Luiz Gonzaga de Castro Júnior é graduado em Administração (1991) e Mestre (1995) pela Universidade Federal de Lavras (UFLA), Lavras, Minas Gerais, Brasil. Doutor em Economia Aplicada pela Universidade de São Paulo, SP, Brasil em 1999. Atualmente é Professor da Universidade Federal de Lavras (UFLA), docente permanente do PPGA/UFLA, orientador de mestrado e doutorado. Coordenador do Centro de Inteligência em Mercados (CIM), Pesquisador líder do Bureau de Inteligência Competitiva do Café, Coordenador de projetos vinculados ao Instituto Nacional de Ciência e Tecnologia do Café (INCT-Café) e ao Polo de Excelência do Café (PEC Secretaria de Ciência e Tecnologia de Minas Gerais). Áreas de atuação: inteligência competitiva, comercialização, mercados de derivativos e gestão de custo.

André Luiz Zambalde é graduado em Engenharia de Telecomunicações pelo Intituto Nacional de Telecomunicações (INATEL), Santa Rita do Sapucaí, MG, Brasil em 1984. Pós-graduado em Administração pela Universidade Federal de Lavras (UFLA), Lavras, MG, Brasil, em 1989, Mestre em Eletrônica pela Universidade Federal de Itajubá (UNIFEI), Itajubá, MG, Brasil em 1991, Doutor em Engenharia de Sistemas e Computação pela COPPE/UFRJ Rio de Janeiro, RJ, Brasil em 2000, possui Pós-Doutorado em Ciência da Computação pela Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, MG, Brasil, em 2005 e Pós-Doutorado em Sistemas e Tecnologias de Informação pela ISEGI-UNINOVA, Portugal, em 2010. Atualmente é professor da Universidade Federal de Lavras (UFLA) nos Departamentos de Computação e Administração. Pesquisador nas áreas de Universidade e Inovação, Gestão do Conhecimento e Inovação, Sistemas de Informação e Informática Aplicada, Estratégia e Governança de Negócios e TI.