PAULO DE OLIVEIRA LIMA JÚNIOR

INTELIGÊNCIA COMPETITIVA NA CAFEICULTURA: MINERAÇÃO TEXTUAL EM NOTÍCIAS PUBLICADAS NA *WEB*

LAVRAS – MG

2016

PAULO DE OLIVEIRA LIMA JÚNIOR

INTELIGÊNCIA COMPETITIVA NA CAFEICULTURA: MINERAÇÃO TEXTUAL EM NOTÍCIAS PUBLICADAS NA *WEB*

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Administração, área de concentração em Gestão de Negócios, Economia e Mercados, para a obtenção do titulo de Doutor.

Dr. Luiz Gonzaga de Castro Júnior
Orientador

Dr. André Luiz Zambalde
Coorientador

LAVRAS – MG

2016

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Lima Júnior, Paulo de Oliveira.
Inteligência competitiva na cafeicultura: mineração textual em notícias publicadas na *web* / Paulo de Oliveira Lima Júnior. – Lavras : UFLA, 2016.
221 p. : il.

Tese(doutorado)–Universidade Federal de Lavras, 2016.
Orientador: Luiz Gonzaga de Castro Júnior.
Bibliografia.

1. Inteligência Competitiva. 2. Mineração Textual. 3. Mercado de Café. I. Universidade Federal de Lavras. II. Título.

PAULO DE OLIVEIRA LIMA JÚNIOR

INTELIGÊNCIA COMPETITIVA NA CAFEICULTURA: MINERAÇÃO TEXTUAL EM NOTÍCIAS PUBLICADAS NA *WEB*

COMPETITIVE INTELIGENCE IN COFFEECULTURE: TEXT MINING OF NEWS PUBLISHED ON THE WEB

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Administração, área de concentração em Gestão de Negócios, Economia e Mercados, para a obtenção do titulo de Doutor.

APROVADA em 26 de julho de 2016.

Prof. Dr. Luciel Henrique de Oliveira	Fundação Getúlio Vargas - São Paulo
Prof. Dr. George Leal Jamil	IETEC
Prof. Dr. Francisval de Melo Carvalho	UFLA
Prof. Dr. Joel Yutaka Sugano	UFLA

Dr. Luiz Gonzaga de Castro Júnior
Orientador

Dr. André Luiz Zambalde
Coorientador

LAVRAS – MG

2016

A Renata, Pedro e Gabriel pela presença e apoio para realização deste trabalho,

Aos meus pais, pelo exemplo de vida,

Dedico

AGRADECIMENTOS

Ao meu pai, pelo exemplo, e minha mãe, fonte de amor, força e determinação tão intensa que ecoará em minha vida para sempre.

A minha esposa Renata e meus filhos Pedro e Gabriel: inspiração e motivação para a vida e carreira.

"O maior risco é não se arriscar" Jorge Paulo Lemann – Obrigado Prof. Luiz Gonzaga pela oportunidade, confiança e orientação.

Ao Prof. André Zambalde pelas contribuições como co-orientador no doutorado e como meu professor no curso de Programação de Computadores em 1985 e 1986.

Ao meu avó Júlio que plantou a semente ao garantir, juntamente com minha mãe, em 1985, minha participação no curso de Programação de Computadores, primeiro contato com a Computação.

A Universidade Federal de Lavras (UFLA) pela oportunidade.

Ao Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) por viabilizar esta jornada.

Ao DAE, funcionários, colegas de curso e professores, especialmente Prof. Mozar José de Brito, Profa. Monica Carvalho Alves Cappelle, Profa. Ana Alice Vilas Boas, Prof. Ricardo Pereira Reis, Prof. Luiz Marcelo Antonialli e Prof. Antônio Carlos dos Santos, por ampliarem meu horizonte nas disciplinas do curso.

Aos professores que contribuíram durante o curso: Prof. Ahmed Esmin, Profa. Cristina Lelis Leal Calegario, Prof. Marcelo Romaniello, Prof. Joel Yutaka Sugano, Prof. Francisval de Melo Carvalho, Prof. Luciel Henrique de Oliveira e Prof. George Leal Jamil.

Aos amigos do Centro de Inteligência em Mercados (CIM), sem os quais a conclusão do trabalho não seria possível: Caio Chain, Diego Humberto

Oliveira, Eduardo César Muniz, Fabrício Guimarães, Murilo Pisciotto e Daniel Basaglia.

Aos amigos Prof. Álvaro Rodrigues Pereira Júnior e Felipe Melo pelas contribuições no início do projeto.

Aos amigos Renegados, presença virtual: Dr. Evandro Faria, Dr. Fabiano Ruosso, Guilherme Ceschiatti B. Moreira, Isac Costa, Jean Melo, J. H. Gomes, Luciano Nachif, Michel Gomes Ank, Rodrigo M. Vaz, Roger Batté e Renato Lenz.

Amigos que contribuíram direta e indiretamente com conversas produtivas sobre o mercado: Leonardo Barros de Oliveira, Dr. Marcelo Gadben, Dr. Arllan Alencar, Prof. Alexandre Pimenta e Bernardo Coelho Vidigal.

RESUMO

A cafeicultura tem um papel significativo para o agronegócio no Brasil, mas é uma atividade de risco elevado pela variação de preço do café que causa impactos em diferentes setores de sua cadeia produtiva. Isto exige dos agentes Inteligência Competitiva para monitorar o ambiente competitivo por meio de um processo contínuo e sistemático de coleta e análise de informações para tomada de decisões em gestão de risco. Notícias com informações que influenciam o mercado de café e afetam a dinâmica da sua cadeia produtiva são publicadas na web diariamente. Entretanto, lidar com o volume e velocidade dessas informações não é uma tarefa trivial, consome recursos humanos, tempo e restringe a análise à capacidade de busca e leitura dos especialistas. E a automatização do processo, apesar do avanço da tecnologia, esbarra em obstáculos no campo sintático como ruídos nos dados e semântico como ambiguidade da linguagem e ausência de contexto. Neste cenário, por meio do processo iterativo do método *Design Science Research*, foi possível, juntamente com especialistas, adquirir conhecimento sobre requisitos de inteligência para cafeicultura e construir artefatos para coletar e classificar automaticamente, pela perspectiva de IC, notícias da web sobre eventos que impactam o mercado de café. Uma avaliação estatística mostrou correlação entre a ocorrência cronológica destes eventos e a série de preço e volatilidade do café, enquanto uma avaliação qualitativa por especialista apontou a relevância das notícias para análise de requisitos de inteligência na cafeicultura. Estes resultados apontam viabilidade de um indicador de evidências qualitativas vindas da *web* que, a saber, sua influência e erro, e confrontado com uma análise qualitativa, permita perceber aumento de volatilidade e viés para delinear um cenário de tomada de decisões para gestão de risco. Desta forma, a pesquisa corrobora a possiblidade de promover Inteligência Competitiva para apoiar decisões sobre gerenciamento de risco e competitividade na cafeicultura por meio de Mineração Textual em notícias publicadas na web.

Palavras – chave: Inteligência Competitiva. Mineração Textual. Mercado de Café.

ABSTRACT

Coffee production plays a significant role in Brazilian agribusiness. However, it is a high-risk activity given the impacts in different sectors of the production chain caused by coffee price variation. This demands Competitive Intelligence for the agents to monitor the competitive environment by means of a continuous and systematic process of information gathering and analysis for the decision-making in risk management. News with information that influence the coffee market and that affect the dynamics of its production chain are daily published online. However, dealing with the volume and speed of this information is not an easy task. It consumes human resources, time and restricts the capacity analysis of the specialists seeking and reading. The automation of the process, despite the advance in technology, meets obstacles in the syntactic field, such as residue on the data, and semantics, such as language ambiguity and absence of context. In this scenery, it was possible to acquire knowledge with the specialists regarding the intelligence for coffee production, by means of the iterative process of the Design Science Research method, and construct artifacts to automatically collect and classify, in the Competitive Intelligence perspective, internet news on events that influence the coffee market. A statistical evaluation showed correlation between the chronologic occurrence of these events and the price series and coffee volatility, while a qualitative evaluation made by specialist pointed the relevance of the news for analyzing the intelligence requisites in coffee production. These results point to the viability of an indicator for qualitative evidence derived from the internet, and its influence and error, while confronting the qualitative analysis, allowing us to perceive an increase in volatility and bias to design a decision-making scenery for risk management. Thus, this research corroborates the possibility of promoting Competitive Intelligence to support decisions regarding risk management and competitiveness in coffee production by means of Text Mining in news published in the internet.

Keywords: Competitive Intelligence. Text Mining. Coffee Market.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

As transformações sociais e culturais, impulsionadas pela revolução tecnológica e da informação na última década, exige adaptação rápida e contínua das organizações, seus processos e tecnologias. Em um ambiente dinâmico e competitivo, os sistemas devem descobrir padrões e relacionamentos em grandes quantidades de dados para identificar oportunidades, aperfeiçoar a tomada de decisões e antecipar as mudanças.

É necessário delinear cenários para previsões de mercado, interpretação de eventos, comportamento dos agentes e tendências. Nesse contexto, a Inteligência Competitiva (IC) tem um papel importante, definida como um conjunto de técnicas e ferramentas que oferecem soluções para transformar dados em informação e conhecimento, com o intuito de monitorar o ambiente competitivo e apoiar a tomada de decisões por meio de um processo contínuo e sistemático de coleta e análise de informações (HOHHOF, 1994; BOSE, 2008; PORTER, 2008; FLEISHER; BENSOUSSAN, 2014).

Atualmente, com o crescimento exponencial da *web*, é possível encontrar um volume significativo de informação textual pública sobre o ambiente competitivo e os concorrentes, que podem emitir sinais estratégicos e influenciar decisões de mercado (CHUNG, 2014) mesmo que suas implicações financeiras não sejam conhecidas de imediato.

Entretanto, textos são dados não estruturados – não estão previamente organizados em um modelo que facilite sua recuperação como em um banco de dados. Para processar textos, os sistemas devem combinar metodologias de análise quantitativa e qualitativa, além de recursos computacionais para aplicação em escala.

Uma abordagem qualitativa para tratamento sistemático de texto é a Análise de Conteúdo (MAYRING, 2000; BARDIN, 2006), que aliada ao uso de *softwares* para análise textual automática apoiada por computador (CATA –

Computer-assisted Automatic Text Analysis), possibilita o uso de grandes quantidades de dados. Combinada as abordagens quantitativas como Mineração Textual (MINER, 2012), a Análise de Conteúdo evolui do campo de ocorrência e frequência de palavras para a inclusão de informação contextual, importante na perspectiva qualitativa para configurar significado (WIEDEMANN, 2013).

Mineração Textual é uma área interdisciplinar com fundamentos teóricos da Estatística, Ciência da Computação e Inteligência Artificial que 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. Ferramentas de IC baseadas em Mineração Textual melhoram a eficiência das organizações para coleta, processamento e análise de informação (BOSE, 2008; CHEN; CHAU; ZENG, 2002).

O desenvolvimento dessas técnicas computacionais de recuperação, processamento e análise possibilitam o uso da informação textual para tomada de decisões em diferentes contextos, como mercado financeiro (LEE; WU; CHEN, 2012; LI et al., 2009; SCHUMAKER et al., 2012), monitoramento de epidemias (COLLIER et al., 2008), lançamento e revisão de produtos (SU, ZHENG; SWEN, 2008; XU et al., 2011; YU et al., 2012) e política (MALOUF; MULLEN, 2008; JUNQUÉ DE FORTUNY et al., 2012).

Uma questão proeminente é, então: como promover, de forma efetiva e eficiente, Inteligência Competitiva em textos, para apoiar decisões estratégicas? a resposta a esta questão se torna ainda mais relevante à medida que aumenta a publicação de notícias na *web* com evidências qualitativas que podem influenciar reações em mercados específicos.

Como pode ser visto em Nunes, Saes e Brando (2004), Abreu et al. (2013) e Martins (2015) um dos mercados influenciados por eventos e fatos relevantes divulgados na mídia é o mercado de café, e a compreensão de

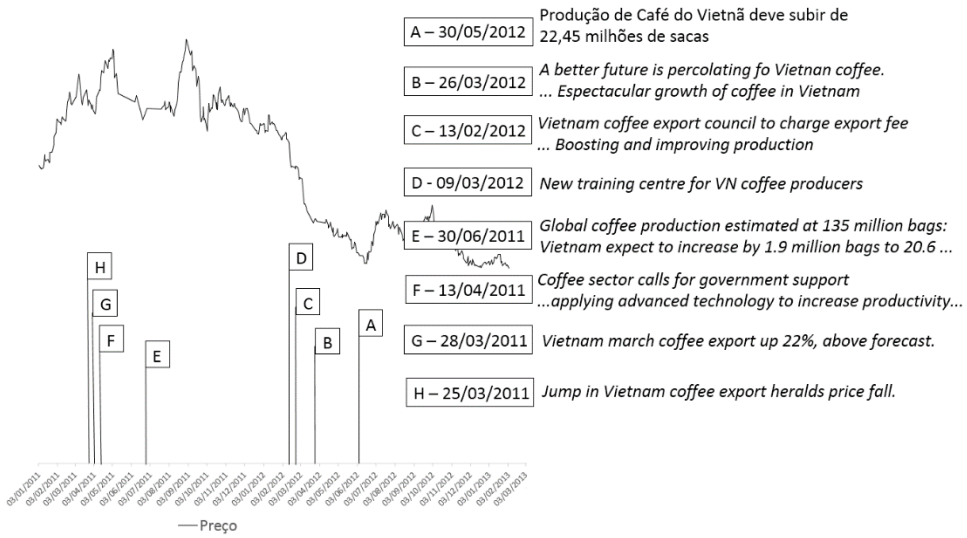
eventos que provocam volatilidade no preço do café e afetam a dinâmica da sua cadeia produtiva são questões complexas e relevantes para a tomada de decisões.

O café é uma das principais *commodities* do Brasil. A cafeicultura gera milhões de empregos, é importante para a economia do país, pois tem papel significativo para o agronegócio, setor expressivo para a composição do Produto Interno Bruto (PIB). Todavia, é uma atividade de risco elevado, pois variações na oferta e demanda e no preço do café causam impactos em diferentes setores de sua cadeia produtiva; afetam decisões sobre estratégias para aumentar competitividade das empresas, projetos de subsídios do governo e programas de incentivo. Isso exige agilidade para tomada de decisões e ajustes ao mercado.

A análise de eventos é uma das tarefas para identificar tendências, estimativas de produção e consumo entre outros fundamentos do mercado de café. Assim, à medida que aumenta o número de notícias publicadas na *web* sobre o mercado de café, a possibilidade de extrair Inteligência Competitiva de textos torna-se relevante na cafeicultura (ABREU et al. 2013).

Como exemplo, a Figura 1 ilustra a variação de preço do café futuro na BM&F no período de 2011 a 2012 (linha cheia).

Figura 1 - Variação do preço do café e ocorrência de notícias.



É possível observar uma queda acentuada no preço no primeiro semestre de 2012, e entre os fatores que influenciam este tipo de movimento está o aumento de oferta mundial de café na safra 2011/2012 para a qual o Vietnã contribuiu com um aumento significativo em sua produção[1], fato publicado pelo *site* Globo Rural[2] em 30/05/2012 – notícia identificada pelo rótulo A na Figura 1. Este fato isolado pode não ser relevante, mas, em conjunto com mais fatores que estimulam a oferta, contribui para delineamento de um cenário que configura impactos no mercado.

Apesar de ter sido um fato inesperado para parte dos agentes do mercado, as notícias sinalizavam expectativa de aumento da oferta pelo Vietnã, com base em eventos passados que indicavam incentivos para aumento da produtividade e exportação, como investimento do governo em tecnologia e

[1] http://www.ico.org/historical/2010-19/PDF/TOTPRODUCTION.pdf
[2] http://revistagloborural.globo.com/Revista/Common/0,,ERT307466-18533,00.html

treinamentos, notícias C, D e F, além de notícias que já indicavam a expectativa de aumento: B, E, G e H na Figura 1.

Entretanto, as notícias sobre o Vietnã destacadas no exemplo e outras com informações que influenciam o mercado de café e afetam a dinâmica da sua cadeia produtiva estão em meio a milhões de outras que são publicadas 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. E a automatização do processo, apesar do avanço da tecnologia, esbarra em obstáculos no campo sintático como ruídos nos dados, e semântico como ambiguidade da linguagem e ausência de contexto. Além dos desafios tecnológicos, há ainda questões gerenciais para o domínio da cafeicultura.

Enfim, este cenário motiva estudos com abordagens quantitativas, qualitativas e tecnologias para Inteligência Competitiva na Cafeicultura a partir de textos publicados na *web*.

A Inteligência Competitiva (IC) é um dos campos de aplicação da Mineração Textual (GUPTA; LEHAL, 2009). Trabalhos sobre o assunto apresentam modelos e sistemas de propósito geral para IC ou aplicados a domínios específicos (ANICA-POPA; CUCUI, 2009; DAI et al., 2013; CHUNG, 2014; YANG; YE, 2014). No mercado de *commodities*, os trabalhos que exploram Mineração Textual em notícias têm foco em petróleo e ouro (FEUERRIEGEL; NEUMANN, 2013).

Porém, pesquisas qualitativas no mercado de café com interpretação de eventos, comportamento dos agentes e tendências (VALKILA, 2014; DUBE; VARGAS, 2013; HEUMESSER; STARITZ, 2013) não têm foco explícito no potencial de IC a partir de evidências qualitativas em textos para a cafeicultura e não exploram a utilização de Mineração Textual.

Em Abreu et al. (2013), IC é usada para identificar tendências para a produção mundial de café a partir de notícias publicadas na *web* por fontes especializadas, porém não se utiliza Mineração *Web* e Textual. Neste sentido o volume de informação analisado se restringe à capacidade dos especialistas.

Neste contexto, entende-se a necessidade de automatização do processo no contexto da cafeicultura, ou seja, adquirir, juntamente com especialistas, conhecimento sobre inteligência para cafeicultura e construir um artefato pela integração de diferentes tecnologias para coletar e classificar automaticamente notícias da *web* e avaliar, com simulações, possíveis impactos e influências destas notícias que contribuam para Inteligência Competitiva na Cafeicultura.

1.1 Problema de pesquisa

Na cafeicultura, para produtores, exportadores e consumidores, a variação de preço no mercado físico e futuro afeta diretamente decisões de comercialização associadas a gerenciamento de risco (BARROS; AGUIAR, 2005; MÓL, 2008; SANTOS et al., 2012), o que exige atenção às mudanças nos fundamentos da oferta e demanda (OLIVEIRA et al., 2013), ao passo que, para o país, o aumento da competitividade entre os países produtores exige atenção às ações dos principais concorrentes da cafeicultura brasileira. Portanto, o monitoramento do ambiente e concorrentes é uma tarefa estratégica para organizações do setor e para o país.

O mercado de derivativos é um dos mecanismos utilizados pelos agentes para o gerenciamento de risco, bem como proteção contra eventuais perdas diante da volatilidade do preço, conforme Hull (2005), via operações de *hedge*. Entretanto, 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 (NUNES; SAES; BRANDO, 2004) até a

ação de especuladores, diferente dos *hedgers*, que procuram ganhos imediatos com variações diárias e intradiárias no mercado.

Estudos sobre volatilidade (MELO; MATTOS, 2012; MÓL, 2008; SANTOS et al., 2012; NUNES; SAES; BRANDO, 2004; FRY; LAI; RHODES, 2011; ZHANG, 2012) e efetividade do *hedge* (FONTES; CASTRO JUNIOR; AZEVEDO, 2003), (PAVÃO, 2010) apontam que períodos específicos de alta volatilidade são explicados por fatores relacionados ao mercado, tais como: quebra de safra, condições climáticas e adversidades para comercialização do produto. Fatores influenciam oferta e demanda de café que por sua vez impactam a variação de preços do café nos mercados.

Existem também informações textuais em notícias sobre ambiente e concorrentes com eventos e fatos relevantes que influenciam oferta, demanda e preço do café, dentre eles: incentivos de governos em países produtores, parcerias entre setor privado e público em países concorrentes, planos de expansão de cafeterias, eventos climáticos, pragas, doenças e outros.

O trecho de notícia extraída da *web*: "Baixa produção de café arábica na Colômbia pode favorecer exportação brasileira" é uma evidência qualitativa que representa uma oportunidade para o Brasil e ao mesmo tempo um evento que contribui para diminuição da oferta. Mas, representa alguma implicação no preço? Ou ainda, um conjunto de evidências qualitativas tem relação com volatilidade de preço no mercado futuro?

Assim, além dos desafios tecnológicos para coletar e classificar automaticamente grande volume de informações da web, há várias questões gerenciais para o domínio da cafeicultura: Que notícias coletar? Que informações devem ser extraídas? Como organizá-las? Como interpretá-las no contexto da cafeicultura? E ainda, existe alguma relação entre esta informação e o preço? É possível verificar relação entre ocorrência de evidências empíricas e

variação de preço? É possível gerar conhecimento para a cafeicultura a partir de informações sobre ambientes e concorrentes disponíveis na *web*?

No que diz respeito a preço, estas perguntas são divergentes à Hipótese de Mercados Eficientes (HME) (FAMA, 1970), segundo a qual os preços refletem toda informação disponível e não dependem de variáveis não fundamentais, como eventos e notícias. Entretanto, trabalhos em Finanças Comportamentais, estudos de eventos e análise de fatos relevantes extrapolam os pressupostos da HME, principalmente da perfeita racionalidade dos agentes, e, conforme Almeida (2011), buscam incorporar aspectos psicológicos no processo de avaliação e precificação de ativos financeiros, além de investigar a influência de eventos e notícias em diferentes mercados (TETLOCK, 2010; SINGER; DREHER; LASER, 2012; ZHANG; SKIENA, 2010). Esses trabalhos indicam que, além de dados fundamentais, os eventos, fatos e notícias geradas por eles são relevantes para o estudo de variação do preço em mercados.

Na cafeicultura, Melo e Mattos (2012) 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 (NUNES; SAES; BRANDO, 2004) e choques na base, positivos ou negativos, demoram um tempo considerável para se dissiparem (MELO; MATTOS, 2012).

A partir da modelagem estatística da série de retorno diário do café futuro BM&F, Martins (2015) 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 fatores climáticos, reforçando que a volatilidade no mercado futuro é vulnerável a esses fatores.

Nunes, Saes e Brando (2004) ressaltam ainda 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.

Também é vantagem competitiva para o Brasil entender a valorização interna e externa do café para o uso de mecanismos de proteção que mantenham o preço competitivo frente às mudanças de cenário do setor (REGO; PAULA, 2012). Portanto, é relevante identificar e monitorar fatores que representam oportunidade e ameaças para que o Brasil mantenha sua posição de mercado, dada a relevância da cafeicultura no país. Conforme Abreu et al. (2013), é necessário monitorar tendências para manter a competitividade do produto brasileiro no mercado mundial.

Nesse contexto, fundamenta-se a questão de pesquisa especificada no presente trabalho: É possível promover Inteligência Competitiva para apoiar decisões sobre gerenciamento de risco e competitividade na cafeicultura por meio de Mineração Textual de notícias publicadas na *web*?

Fundamentado no modelo de Pellissier e Nenzhelele (2013b) para o processo de IC, as questões específicas de pesquisa foram delimitadas em cada fase para o contexto da cafeicultura:

Planejamento e Direção: Questão 1 – Quais os requisitos de inteligência para gerenciamento de risco e competitividade em cafeicultura?

Coleta de Informações: Questão 2 – Quais informações relevantes devem ser coletadas da *web* para atender os requisitos da Questão 1?

Triagem, captura e armazenamento de informações: Questão 3 – Como estas informações devem ser organizadas?

Análise de Informações: Questão 4 – Por quais modelos a informação deve ser interpretada e analisada para IC?

Difusão de Inteligência: Questão 5 – Como a inteligência deve ser apresentada para apoiar a tomada de decisões?

Existem tecnologias relevantes para apoiar atividades de cada fase do modelo: Recuperação da Informação e *Webmining* para a fase de coleta de informações, Processamento de Linguagem Natural e Mineração Textual para a fase de organização da informação e fase de análise da informação, juntamente com técnicas de IC como: Análise de Forças, Oportunidades, Fraquezas e Ameaças (JOHNSON; SCHOLES; WHITTINGTON, 2008) também denominada Análise SWOT (termo do idioma inglês acrónimo de *Strengths*, *Weaknesses*, *Opportunities*, *Threats*), Forças Competitivas de Porter (PORTER, 2008) e Análise de Eventos.

Essas técnicas são propostas por Dai et al. (2013) para obter IC a partir de textos em um sistema baseado em Mineração Textual para Inteligência Competitiva – TMCIS (*Text Mining based Competitive Intelligence System*). Sua pesquisa apresenta como aplicar tecnologias associadas a IC para decisões estratégicas em um contexto geral. O modelo proposto é referência neste trabalho aplicado para o contexto da cafeicultura.

A opção por esta tecnologia fundamenta-se na possiblidade de examinar automaticamente dados não estruturados ou semiestruturados de textos provenientes da *web*.

1.2 Objetivos

Considerando o problema e as questões apresentadas, O objetivo geral do trabalho foi propor, desenvolver e avaliar o protótipo de um sistema baseado em Mineração Textual para Inteligência Competitiva (TMCIS – *Text Mining based Competitive Intelligence System*) (DAI et al., 2013) aplicado às questões da cafeicultura, como produto do modelo de processo de IC adotado como referência

Para tal, os objetivos específicos pela perspectiva prática foram:

a) Investigar o conceito e a aplicação de TMCIS para apoio à obtenção de Inteligência Competitiva na Cafeicultura;
b) Desenvolver o protótipo de um TMCIS, pesquisar, extrair da *web*, classificar e analisar evidências qualitativas, pré-definidas por especialistas, em notícias sobre o mercado de café;
c) Verificar em que medida a informação qualitativa coletada da *web* apresenta correlação com a variação de preço do café em diferentes mercados – Físico, BM&F Bovespa e Bolsa de Nova York.
d) Avaliar a contribuição do protótipo para apoiar Inteligência Competitiva em uma organização real do setor cafeeiro.

1.3 Contexto da pesquisa

A pesquisa foi realizada em uma organização real, o Centro de Inteligência em Mercados (CIM) da Universidade Federal de Lavras (UFLA). O CIM é um grupo de pesquisa com projetos voltados para a cafeicultura brasileira. Entre os projetos está o Bureau de Inteligência do Café (BIC), iniciado em 2010 com apoio do Polo de Excelência do Café (PEC), da Secretaria de Estado de Ciência Tecnologia e Ensino Superior de Minas Gerais. Posteriormente, com apoio Consórcio de Pesquisa do Café, passou a integrar a

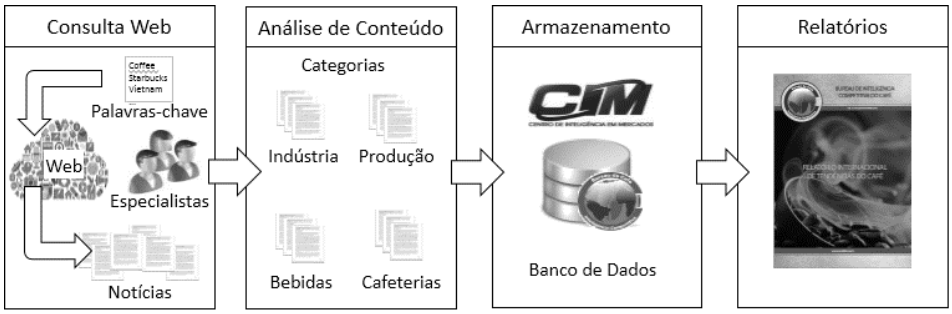
Agência de Inovação do Café (Inovacafé), organização gerenciada pela UFLA com o objetivo de integrar conhecimentos de áreas relacionadas ao café e desenvolver soluções e inovações para o setor.

O BIC tem como objetivo oferecer informações e análises sobre o setor cafeeiro que contribuam para planejamento e tomada de decisões pelos agentes da cadeia agroindustrial do café. Entre seus produtos estão os relatórios:

a) Relatório Internacional de Tendências do Café;
b) Potenciais Concorrentes do Café Brasileiro.

Para redigir os relatórios, profissionais do CIM realizam análises de notícias sobre agentes e setores da cadeia produtiva publicadas na *web*. Uma equipe de especialistas monitora notícias sobre o mercado diariamente com ferramentas de busca. Elas são coletadas, armazenadas em um banco de dados e classificadas de acordo com relevância para determinadas áreas da cadeia produtiva, conforme Figura 2.

Figura 2 - Coleta, classificação, armazenamento e produção de relatório.



Neste ambiente, seguindo a abordagem de observação participante juntamente com os especialistas, foi possível verificar: i. O que é extraído da *web*; ii. Como o resultado deve ser organizado; iii. O que é analisado em busca

de Inteligência Competitiva e; iv. Definir os requisitos para construção e avaliação do sistema.

Para esta pesquisa, foi considerado o banco de dados com 3.098 notícias sobre o mercado de café, em inglês, coletadas pelos especialistas no período de 01/01/2011 a 15/11/2015. Para análise de gerenciamento de risco, foram utilizadas as séries históricas de preços do café arábica origem Sul de Minas, preço do contrato futuro de café negociado na BM&F Bovespa com o código ICF, série de preços do Café na Bolsa de Nova York e série histórica de cotações do dólar BM&F Bovespa para conversão de moeda quando necessário.

1.4 Organização da Tese

O Capítulo 2 apresenta o paradigma metodológico, o restante do texto descreve a pesquisa seguindo a metodologia *Design Science Research:* o Capítulo 3 apresenta a revisão de literatura que direciona etapas do método; o Capítulo 4 apresenta o processo de entendimento do problema e os resultados da interação com especialistas do BIC; o Capítulo 5 apresenta o modelo proposto na etapa de sugestão para o sistema de Inteligência Competitiva para Cafeicultura baseado em Mineração Textual; o Capítulo 6 apresenta o desenvolvimento da arquitetura proposta no Capítulo 5 e os resultados do sistema; o Capítulo 7 apresenta a avaliação; o Capítulo 8 considerações finais e limitações; e o Capítulo 9 apresenta as conclusões e os trabalhos futuros.

2 PARADIGMA METODOLÓGICO – *DESIGN SCIENCE*

Por buscar entendimento sobre fenômenos criados pelo homem (publicação de dados na *web* e variação de preço do café) com o uso de tecnologia (Mineração Textual), para resolução de problema em um contexto específico (gestão de risco na cafeicultura), esta pesquisa segue o paradigma *Design Science* pelo método *Design Science Research*.

Diferente das ciências naturais que pesquisam como se comportam ou interagem classes de objetos e fenômenos naturais (Biologia, Química, Física) e sociais (Economia, Sociologia), as ciências do artificial (SIMON, 1996) lidam com fenômenos produzidos, inventados ou que sofrem intervenção do homem, ou artificiais – computadores, organizações, economia, entre outros.

Para os fenômenos artificiais nem sempre as ciências tradicionais se sustentam em seus objetivos de explorar, descrever ou explicar, pois é necessário buscar como as coisas devem ser para alcançar determinados objetivos, seja para solucionar um problema ou projetar e construir algo que ainda não existe, objetivos das ciências do artificial (LACERDA et al., 2013).

Neste contexto, a Ciência do Projeto ou *Design Science* surge como um corpo de conhecimento rigoroso e validado com o propósito de adquirir conhecimento para a concepção e desenvolvimento de artefatos (AKEN, 2004) que realizem objetivos (SIMON, 1996). Firma-se como paradigma epistemológico para pesquisas orientadas à solução de problemas e ao projeto de artefatos, reforçada pelos principais argumentos encontrados na literatura, apresentados por Dresch (2013):

a) O mundo em que vivemos é mais artificial do que natural, logo, uma ciência que se ocupe do artificial é necessária segundo Simon (1996), Le Moigne (1994) apud Dresch (2013);

b) As ciências tradicionais geram conhecimento sobre coisas que existem, não se ocupam do projeto ou estudo de sistemas que ainda não existem (SIMON, 1996; MARCH; SMITH, 1995; AKEN, 2004; VAN AKEN, 2005);
c) Apenas o entendimento acerca de um problema não é suficiente para resolvê-lo, o conhecimento gerado nas ciências tradicionais é de cunho fortemente exploratório e analítico, não necessariamente contribuindo significativamente para a utilização em situações reais – lacuna entre teoria e prática (ROMME, 2003; VAN AKEN, 2005);
d) A construção adequada do conhecimento deve ocorrer a partir do processo de pesquisa incluindo a interação entre objeto e observador (LE MOIGNE, 1994 apud DRESCH, 2013);

Diante destes argumentos, *Design Science* é a ciência que procura desenvolver e projetar soluções para melhorar sistemas existentes, resolver problemas ou, ainda, criar novos artefatos que contribuam para melhor atuação humana, seja na sociedade, seja nas organizações (DRESCH, 2013). Nesta definição, artefatos são a interface entre o ambiente interno e externo de um determinado sistema (SIMON, 1996) e soluções devem ser viáveis para a realidade em um contexto, não necessariamente ótimas.

Conforme Dresch (2013), a pesquisa orientada por este paradigma se diferencia pelos objetivos, acrescenta ponto de partida e o método abdutivo. Além de reposta para uma questão importante, acrescenta como ponto de partida a solução para um problema prático, ou classe de problemas – constata-se a necessidade de formalizar ou desenvolver um artefato a partir da observação da realidade, em contraponto com as pesquisas tradicionais que se iniciam pela necessidade de compreensão de um fenômeno em profundidade.

Este ponto de partida diferencia também os objetivos que são: prescrever ou projetar. Assim o conhecimento produzido é uma prescrição para resolver um problema real ou um projeto para um novo artefato, ambos com validade acadêmica e pragmática – úteis para os profissionais interessados.

Acrescenta o método científico abdutivo em conjunto com os demais, que consiste em estudar fatos e propor uma teoria para explicá-los, ou levantar hipóteses explicativas para fenômenos ou situações. O método abdutivo sugere o que pode ser, e não pretende afirmar o que é ou o que deva ser.

A pesquisa modelada neste quadro epistemológico tem como problema de ordem prática a necessidade de Inteligência Competitiva para gestão de risco na cafeicultura e como artefato um sistema baseado em Mineração Textual para desenvolver IC a partir de notícias publicadas na *web* – fenômenos artificiais.

Do ponto de vista prático, existem ferramentas para Mineração Textual que podem ser utilizadas na construção de um sistema, entretanto, a revisão da literatura não revela a existência de bases de treinamento ou dicionários léxicos específicos para a cafeicultura que possibilitem a aplicação da tecnologia de Mineração Textual por aprendizado de máquina para o domínio do problema. A necessidade de criar estes mecanismos justifica a adoção do paradigma.

O método de pesquisa que operacionaliza os conceitos de *Design Science* é o *Design Science Research*, detalhado na próxima seção como etapas deste trabalho.

2.1 *Design Science Research*

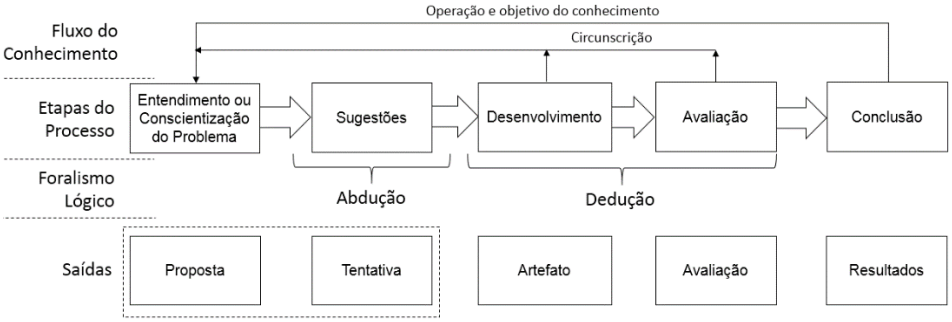
A *Design Science Research* tem como objetivo estudar, pesquisar e investigar o artificial e seu comportamento sob a perspectiva acadêmica e organizacional (BAYAZIT, 2004). É o método que operacionaliza e fundamenta a condução da pesquisa quando o resultado é um artefato ou prescrição.

A ideia central de *Design Science Research* é que a aquisição de conhecimento e a solução de um problema acontecem pela construção e aplicação de um artefato. Os artefatos podem ser constructos, modelos, métodos, instanciações e aprimoramento de teorias existentes, e a proposta de solução deve ser passível de generalização para uma classe de problemas.

Para Vaishnavi e Kuechler (2004), *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. Machado et al. (2013) ressaltam que a aplicação deste método é relevante para a Administração, pois gera tecnologia em gestão, procedimentos, metodologias e soluções para a resolução de problemas práticos da gestão. O trabalho de Sordi, Meireles e Sanches (2011) mostra que este método como abordagem de pesquisa vem crescendo na área graças ao caráter aplicado da Administração e apresenta um número representativo de pesquisas publicadas pela Academia Brasileira de Administração.

Vários autores formalizaram métodos para operacionalizar pesquisas no paradigma *Design Science* em diferentes áreas (DRESCH, 2013). Nesta pesquisa o modelo adotado é adaptado do proposto por Vaishnavi e Kuechler (2004), aperfeiçoado de Takeda, Veerkamp e Yoshikawa (1990), com contribuições de Manson (2006), conforme Figura 3, por ter suas raízes em Sistemas de Informação e prever a interação entre todos os envolvidos na pesquisa para que entendam e aprendam com o processo de construção do artefato.

Figura 3 - Etapas do método *Design Science Research*.



Fonte: Adaptado de Takeda, Veerkamp e Yoshikawa (1990), Vaishnavi e Kuechler (2004) e Manson (2006).

A investigação tem início pelo conhecimento de um problema ou oportunidade de pesquisa na etapa de Entendimento ou Conscientização do Problema. Deve-se identificar e compreender o problema a ser solucionado e definir qual desempenho é 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 (MANSON, 2006).

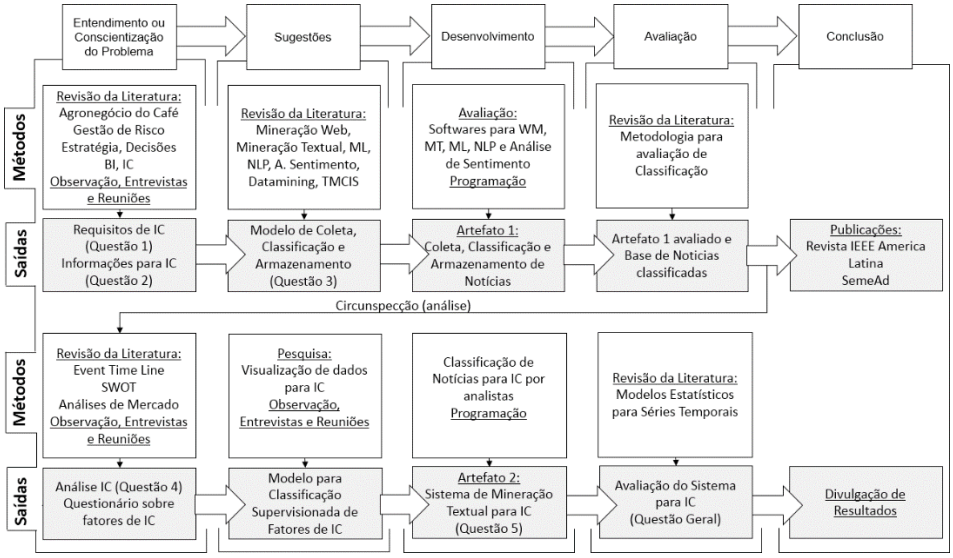
Na etapa de sugestão, são elaborados um ou mais modelos de tentativa para a solução do problema a partir da existência de conhecimento/teoria de base sobre o problema. Predomina o método abdutivo que exige processo criativo e conhecimento prévio para proposta de soluções (VAISHNAVI; KUECHLER, 2004). 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, os adequados são posteriormente avaliados na quarta etapa. Caso o artefato seja inadequado durante as etapas de desenvolvimento ou avaliação, retorna-se à primeira etapa - revisão do entendimento do problema. Este ciclo denominado circunscrição (circunspecção) 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 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 em duas fases divididas por uma circunspecção, conforme Figura 4.

Figura 4 - Etapas de *Design Science Research* na pesquisa.



Na primeira fase e primeira etapa do método, o problema foi delineado por meio de revisão da literatura sobre o Agronegócio do Café e gestão de risco com derivativos, revisão da literatura sobre os conceitos de Estratégia, Tomada de Decisões, *Business Intelligence* e Inteligência Competitiva que guiaram a análise de dados existentes, entrevistas e reuniões formais com os profissionais, das quais foram definidos os requisitos de IC para a cafeicultura e quais

informações seriam coletadas da *web* para atender esses requisitos, respondendo as questões propostas 1 e 2.

Na etapa de sugestões, a partir de revisão da literatura sobre *Datamining*, Mineração Textual e Análise de Sentimento em paralelo a estudo de modelos de TMCIS em Dai, Kakkonen e Sutinen (2010, 2011a, 2011b) e Dai et al. (2013), foi proposto um modelo para coleta, classificação automática e armazenamento de notícias para o mercado de café de acordo com os requisitos levantados com os especialistas do BIC, em resposta à questão 3.

Na etapa de desenvolvimento, constatou-se a necessidade de utilização de diferentes tecnologias para materializar uma instância do modelo proposto na etapa anterior, o que levou a experimentos com ferramentas disponíveis para coleta de dados na *web*, Mineração Textual e Processamento de Linguagem Natural, paralelamente ao desenvolvimento do artefato.

Por se tratar de um *software*, para materializar o artefato proposto, foi realizado o projeto da arquitetura, integração das tecnologias e codificação do sistema no ambiente do BIC por uma adaptação da metodologia ágil de desenvolvimento de *software* denominada XP – *eXtreme Programming* (BECK, 2000) por seus princípios básicos pertinentes a esta pesquisa: *feedback* rápido, presumir simplicidade e mudanças incrementais. Desta forma, o desenvolvimento acontece pela aplicação iterativa das tarefas: codificar, testar, ouvir usuário, refatorar.

A atividade central do artefato foi coletar e classificar notícias da *web* em categorias da cafeicultura definidas pelos especialistas do BIC. Para tal, foram utilizados algoritmos de aprendizado de máquina com uma base de treinamento criada a partir do banco de dados rotulados do BIC. Portanto, o artefato foi validado de acordo com os parâmetros observados para a tarefa de categorização textual supervisionada propostos por Sebastiani (2002) e Baeza-

Yates e Ribeiro-Neto (2013). Os resultados foram divulgados em forma de artigos.

A conclusão da etapa de avaliação apontou para a necessidade de retorno à primeira etapa do método em busca de novo entendimento do problema quanto a ferramentas de análise para IC, impulsionando a revisão da literatura em *SWOT*, *Event Timeline* e Forças Competitivas de Porter. Ao mesmo tempo, entrevistas e reuniões formais com os especialistas resultaram na definição de novas categorias para classificação quanto a fatores específicos para IC na cafeicultura definindo um modelo de análise para interpretação das notícias coletadas da *web* pelo artefato gerado na fase anterior.

A sugestão seguinte, em conjunto com os especialistas, foi um modelo de melhoria do artefato com inclusão de um módulo para Análise de Conteúdo ou Módulo de Supervisão, gerando desta forma uma base de treinamento para desenvolver IC a partir de notícias coletadas pelo artefato. Na etapa de desenvolvimento, após a codificação do módulo, os especialistas classificaram as novas notícias, criando uma base de dados rotulada possibilitando classificação automática e análise de acordo com as novas categorias voltadas para IC.

A etapa de avaliação utiliza critérios quantitativos para verificar o desempenho de classificação textual por aprendizado de máquina, descritos por Sebastiani (2002) e Baeza-Yates e Ribeiro-Neto (2013): estudo descritivo de correlação e modelo estatístico ARIMA (BOX et al., 2015) para verificar a relação entre os dados coletados da *web* com o preço do café e a volatilidade, e critérios qualitativos para avaliar o valor do resultado do protótipo para a aquisição de IC pelos especialistas do BIC. Essa etapa é delineada e divulgada neste trabalho, bem como as demais, conforme organização descrita na Seção 1.4.

3 REFERENCIAL TEÓRICO

Nesta seção, são apresentados trabalhos relevantes com contribuições teóricas e práticas em áreas específicas que constituem fundamentos para o desenvolvimento desta pesquisa. Inicialmente são apresentadas características do agronegócio do café no Brasil e a importância do uso de gerenciamento de risco e informação para tomada de decisão neste setor. Para contextualizar Inteligência Competitiva nesta pesquisa, são introduzidos os conceitos de Modelo de Negócios, Estratégia, Tomada de Decisões e *Business Intelligence* (BI). Em seguida são apresentados os temas: Mineração de Dados, Análise de Conteúdo, Mineração Textual e Análise de Sentimento que formam a base tecnologia deste projeto. Ao fim da seção são apresentados trabalhos que aplicam Mineração Textual em mercados, contextos relacionados e ferramentas.

3.1 O agronegócio café

O Brasil é responsável por quase um terço da produção mundial de café. Segundo dados da *ICO – International Coffee Organization*, na safra 2015/2016, a cafeicultura brasileira teve produção total de 43.235.000 sacas de um total mundial de 143.371.000 de sacas. O café tem papel significativo para o agronegócio – de janeiro a dezembro de 2015 representou 7% das exportações (Ministério da Agricultura, 2016) – que é expressivo para a composição do Produto Interno Bruto (PIB) do Brasil. É importante para a economia do país, gera milhões de empregos, além de ter a produção cafeeira como fixadora de mão de obra no campo com responsabilidade social.

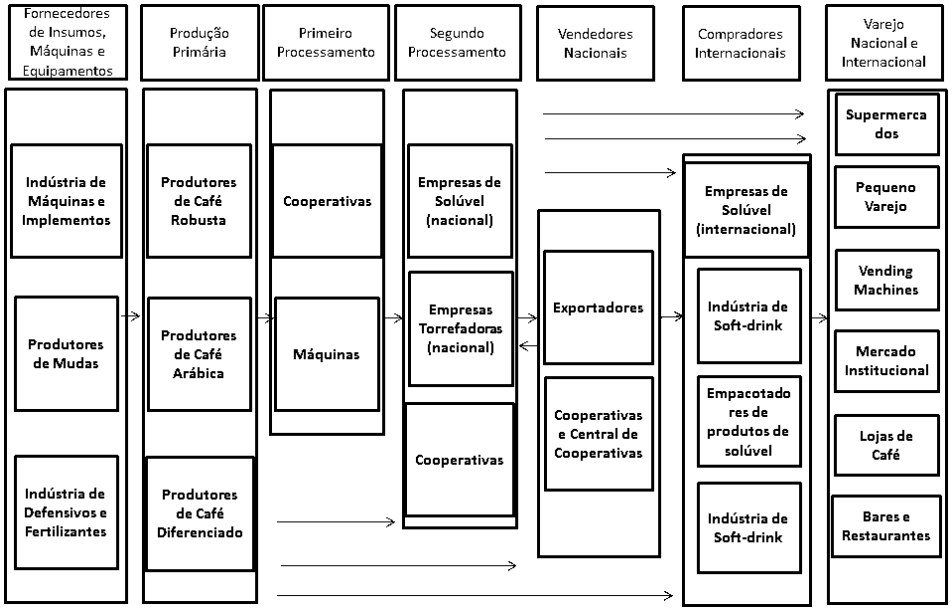
Como setor estratégico, possui o Fundo de Defesa da Economia Cafeeira (Funcafé) criado pelo Decreto-Lei nº 2.295/86 e estruturado pelo Decreto nº 94.874/87 para pesquisas, incentivo à produtividade e competitividade, qualificação de mão de obra, publicidade, linhas de financiamento para custeio, colheita, estocagem e aquisição de café.

Em âmbito internacional, o Brasil ocupa a posição de maior produtor e exportador mundial há alguns anos e o segundo maior consumidor em 2015, conforme dados da ICO – *International Coffee Organization*, que também aponta que alguns países começaram a se destacar como produtores na última década como Vietnã e Indonésia, mas o título permanece com o Brasil, o que reforça a importância da *commodity* na balança comercial. No ano-safra 2015/2016 o país registrou exportação recorde de 36,89 milhões de sacas de café, conforme levantamento do Conselho dos Exportadores de Café, apesar disto a receita cambial caiu 7% em relação à última receita, principalmente pela queda dos preços internacionais do produto.

O Brasil tem vantagens em relação a outros países produtores pelo parque cafeeiro complexo e diverso – existem duas variedades principais: arábica (*Coffea Arabica*) e robusta ou conilon (*Coffea Canephora Pierre*) – e pela tecnologia: fertilização irrigação e mecanização (NAKAZONE; SAES, 2004). Entretanto é necessário entender a dinâmica de valorização interna e externa do café, identificar e monitorar tendências e fatores que representam oportunidades e ameaças para que o Brasil mantenha sua posição de mercado, aloque adequadamente recursos e direcione políticas públicas.

A cadeia produtiva do café no Brasil é representada pelo Sistema Agroindustrial, conforme apontado pelo estudo de Farina e Zylbersztajn (1998). Cadeia produtiva ou cadeia de suprimento é um termo usado para indicar uma sequência de estágio de materiais e processos para a fabricação de produtos e serviços. A Figura 5 mostra a cadeia produtiva do café no Brasil dividida em vários segmentos que se relacionam pelas transações.

Figura 5 - Sistema Agroindustrial do Café no Brasil.



Fonte: (FARINA; ZYLBERSZTAJN, 1998).

Cada segmento possui agentes ou novos segmentos complexos, por exemplo, o segmento Fornecedores de Insumos, Máquinas e Equipamentos tem a Indústria de Máquinas e Implementos que por si só constituem uma cadeia.

A produção possui parque cafeeiro estimado em 2,25 milhões de hectares, 287 mil produtores, maioria mini e pequenos, distribuídos em aproximadamente 1.900 municípios (BRASIL, [2016?]), participando de associações e cooperativas nas diferentes dimensões da cadeia.

Ao longo dos anos, o setor passou por mudanças estruturais que trouxeram a necessidade de rever as inter-relações na cadeia produtiva. Mudanças geográficas – localização da lavoura e distribuição geográfica da produção, mudanças políticas – políticas setoriais e variação no papel do governo (CARVALHO, 2002), a ampliação dos mercados internacionais impacta a capacidade de governos em implementar programas autônomos e

dificulta medidas intervencionistas como fixação de preço pelo governo, o que inibe o setor público e impulsiona o setor privado (FILENI, 1999), e mudanças tecnológicas – o setor passou a utilizar sistemas de produção inovadores, diferenciação de mercado pela qualidade, redução de custos via elevação de produtividade e adoção que novas tecnologias de produção pré e pós-colheita (MARTIN; VEGRO; MORICOCHI, 1995; GROSSI, 1998 apud CARVALHO, 2002), irrigação (SANTINATO; FERNANDES, 2000) e cultivares (CARVALHO et al., 2011).

Estas mudanças contribuíram para aumentar a competitividade do setor e por consequência a dinâmica do mercado. Além desta dinâmica, aspectos climáticos – precipitação, vento, temperatura, umidade do ar – pragas e doenças impactam direta e indiretamente a cafeicultura. Outro fator de impacto é o comportamento do preço do café no mercado físico e futuro que sofre períodos de alta volatilidade influenciado por fatores de mercado, climáticos e macroeconômicos: taxa de juros, câmbio e política monetária entre outros. A soma destes fatores caracteriza a cafeicultura como uma atividade de risco elevado que exige agilidade dos agentes da cadeia produtiva para tomada de decisões e ajustes ao mercado.

3.2 Gerenciamento de Risco

Segundo Marques, Mello e Martines Filho (2006), nos mercados agrícolas há o risco de produção (produto errado, perda de produção, impropriedade do produto às necessidades do cliente, etc.) para o qual os produtores tradicionalmente utilizam os seguros agrícolas e tecnologias adequadas para evitar os fatores associados à quebra de safras. E o risco de preço, associado a subida ou queda de preço do café no mercado, acarretando prejuízos para produtores e consumidores. Para administrar esse risco, existem

os mercados a termo e de derivativos: opções e contratos futuros, em que é possível fixar preços antecipadamente.

A cadeia produtiva do café inclui agricultores, fornecedores de insumos e serviços, indústrias de processamento e transformação, agentes de distribuição e comercialização e consumidores finais. Quanto à atuação no mercado, pode-se observar duas classes gerais de agentes: os que produzem e vendem o café (produtores) e os que têm o café como insumo ou compram o café (consumidores). Todos expostos à volatilidade do preço.

O risco de preço para os produtores está relacionado ao custo de produção. Em determinado nível, o preço do café não cobre os custos, gera prejuízos e inviabiliza a atividade. Fatores que geram queda no preço do café representam risco nesse caso. A queda afeta também agentes que realizaram compra antecipada para garantir matéria-prima. Os consumidores têm prejuízo ou inviabilidade pela alta no preço, por exemplo, o exportador que fixou o preço de venda com algum recurso do mercado financeiro sem adquirir o café.

Uma vez que há oscilações e incertezas que impedem estimativas precisas de preço, é importante criar vantagens competitivas na produção e comercialização, tanto para crescimento econômico do país como para viabilizar a atividade dos agentes da cadeia produtiva. Órgãos competentes precisam de subsídios para decisões estratégicas em linhas de crédito para custeio, estocagem, aquisição de café, opções e operações no mercado futuro e capital de giro para indústria. Produtores e consumidores por sua vez também devem estar atentos aos fatores que provocam oscilações e reagirem em tempo para administrar o risco de preço.

Conforme Marques, Mello e Martines Filho (2006) um dos principais instrumentos para a aquisição de vantagem competitiva é um mercado de derivativos eficiente e abrangente. Neste é possível administrar o risco na comercialização do café pela negociação de derivativos agrícolas: opções e

contratos futuros. No Brasil, o principal mercado é a Bolsa de Mercadorias e Futuros – BM&F Bovespa.

Por meio de operações com opções ou contratos de futuros agrícolas, é possível diminuir o risco pela estabilização de preços. A operação comum no mercado de café com contratos futuros é denominada *hedge*. Porém, conforme Hull (2005) a operação via *hedge* não é trivial, há incerteza quanto ao momento adequado para abertura e encerramento da operação e são vários eventos que afetam a variação de preço, desde clima até especuladores que buscam ganhos com variações diárias e intradiárias.

Para entender o uso de derivativos para gerenciar risco, a próxima seção descreve em abrangência a dinâmica de formação de preços, estrutura e funcionamento da bolsa e operações de *hedge*.

3.2.1 Gerenciamento de Risco no Mercado de Café: preço

Como *commodity* agrícola, o café é considerado um produto pouco elástico ou inelástico, com preço formado em um sistema de leilão em mercados, sujeito às forças de compradores e vendedores, influenciado por oferta e demanda. Existe o mercado físico e futuro de café. O físico ou disponível é o mercado em que o produto é negociado em dinheiro, o produtor entrega seu produto e recebe o pagamento no ato - à vista ou a prazo (MARQUES; MELLO; MARTINES FILHO, 2006). Há cotações diárias de diferentes tipos e qualidades de café em Minas Gerais (Sul de Minas, Zona da Mata e Cerrado), São Paulo (Mogiana, Pinhal e Garça), Espírito Santo (Vitória e Colatina), Paraná (Norte), Bahia (Vitória da Conquista e Barreiras) e Roraima (Cacoal) (MERCADO, 2014).

Mercados futuros organizados têm como propósito o estabelecimento de contratos futuros, coordenados por bolsas, que visam estabelecer normas e procedimentos de negociação (RIBEIRO; SOUSA; ROGERS, 2006). Contrato

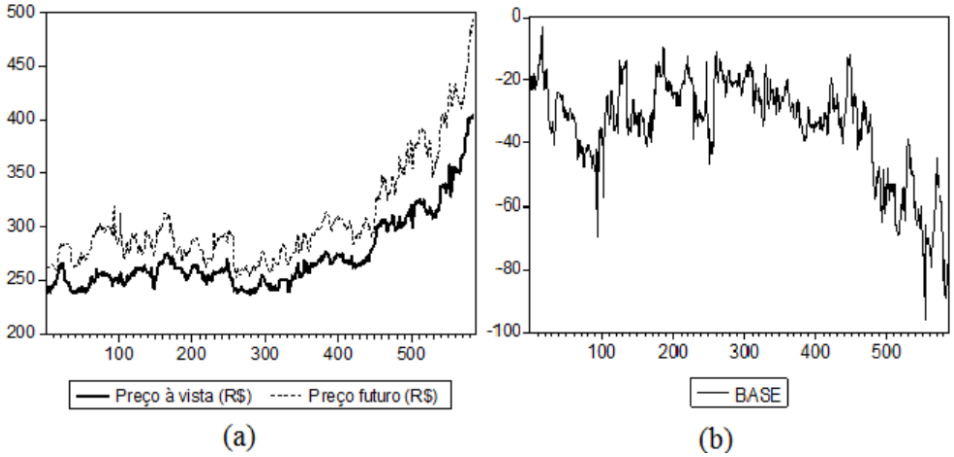
futuro é um derivativo – preços derivam do preço no mercado físico (a vista) – que representa um compromisso de comprar ou vender determinado ativo objeto em uma data específica no futuro, por um preço pré-estabelecido (HULL, 2005). É um mecanismo utilizado para gerenciamento de risco, pois permite fixar, no presente, o preço do ativo para uma data futura, reduzindo a exposição às variações do preço. São ativos objeto: ações, índices e *commodities*.

No Brasil, existem dois contratos futuros de café padronizados e negociados na BM&F Bovespa, o que tem como objeto de negociação o café cru, em grão, de produção brasileira, *coffea* arábica, tipo 4-25 (4/5) ou melhor, bebida dura ou melhor, para entrega no Município de São Paulo (Bovespa, 2015a). E o tipo 6-25 (6/7) (BM&FBOVESPA, 2015b) com as mesmas demais características. Para ambos, o tamanho de um contrato é de 100 sacas de 60 kg líquidos (seis toneladas métricas) cotados em dólares dos EUA, sendo que o tipo 4/5 possui maior liquidez.

As operações com contratos futuros são denominadas *hedge*. No caso do café, o produtor pode se proteger da queda no preço vendendo contratos futuros de café, e o consumidor da alta de preço comprando contratos. Existem outros mecanismos no mercado para administrar risco – mercado a termo, Cédula do Produtor Rural (CPR) e opções.

Ao realizar o *hedge* 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 (HULL, 2005), como mostra a Figura 6 (a) Preço à vista e futuro e (b) a base.

Figura 6 - (a) Preço no mercado à vista e futuro, (b) Base.



Fonte: Melo e Mattos (2012).

O *hedge* perfeito é quando o preço no mercado futuro é igual ao preço no mercado físico local no dia do vencimento do contrato futuro – base igual a zero. Isso dificilmente ocorre (MÓL, 2008), pois, o prêmio pelo risco, taxa de juros, custo de carregamento, transporte e armazenamento, localização da praça, qualidade do produto e características intrínsecas da comercialização interferem na sua formação. A base é volátil, assim ao eliminar o risco de variação de preço passa a assumir o risco da variação da base ou risco de base.

O risco de base é investigado por pesquisas com modelos matemáticos e estatísticos que procuram explicar características da volatilidade do mercado (MELO; MATTOS, 2012; MÓL, 2008; SANTOS et al., 2012; FRY; LAI; RHODES, 2011; ZHANG, 2012) e efetividade do *hedge* (FONTES; CASTRO JÚNIOR; AZEVEDO, 2003; PAVÃO, 2010). Com uma abordagem quantitativa estes estudos apontam que períodos específicos de alta volatilidade são explicados por fatores relacionados ao mercado, quebra de safra, condições climáticas e adversidades para comercialização do produto.

Há sinais de fatos estilizados na maioria das séries de retornos (MÓL, 2008) – aproximação teórica de um fenômeno observado empiricamente. Fatores influenciam oferta e demanda de café que, por sua vez, impactam a variação de preços do café nos mercados (CASTRO JUNIOR, 2008; HULL, 2005), assim a compreensão destes fatores é base para tomada de decisões no cenário de incerteza por diferentes agentes. Além disso, também é importante compreender que o mercado está sujeito à ação de diferentes agentes: *hedgers*, especuladores e arbitradores.

O participante que utiliza o mercado futuro para reduzir e eliminar o risco proveniente de perdas e oscilação nos preços é o *hedger*. Especuladores são aqueles que buscam ganhos com variações diárias e intradiárias – operações que iniciam e terminam no mesmo dia, ou em minutos – são mais ativos durante o pregão e contribuem para a formação do preço e de liquidez do mercado.

Os arbitradores se beneficiam da má formação dos preços, procurando por lucros sem a exposição a riscos, por meio da realização de operações concomitantes em mais de um mercado (MARTINS, 2005). A atuação deles contribui para o equilíbrio de preços no mercado à vista e futuro, por exemplo.

Os agentes diferem quanto aos objetivos, mas independente da estratégia, a tomada de decisão passa por uma tarefa comum: previsão de preço, tema complexo estudado em todo mercado financeiro, discutido na próxima seção.

3.2.2 Análises do mercado

Para tomar decisões sobre *hedge*, especulação ou arbitragem no mercado de café, analistas procuram deduzir a trajetória futura do preço pela compreensão de fatores de oferta e demanda que o afetam. A previsão de preços é um tema complexo, pois questiona a principal sustentação na Teoria de

Finanças Moderna: a Hipótese de Mercados Eficientes – HME (FAMA, 1970). É estudado em todo mercado financeiro: ações, opções e futuros.

A HME afirma que os preços em mercados refletem todas as informações publicamente disponíveis, e um agente não consegue alcançar consistentemente retornos superiores à média do mercado considerando informações públicas disponíveis: preços passados, dados fundamentalistas, notícias e fatos relevantes. O que acontece com a série de preço no mercado é um passeio aleatório (*random walk*), imprevisível. As variações da série são variáveis aleatórias de um processo estocástico denominado ruído branco, caracterizado por média zero, variância constante, independência das informações e $\varepsilon_t = p_t - p_{t-1}$. O preço hoje não depende do de ontem.

A HME considera que as informações estão disponíveis a todos no mesmo instante, expectativas são homogêneas e que todos os agentes são racionais e maximizam sua utilidade esperada em função da sua aversão ao risco. Porém, trabalhos em Finanças Comportamentais (KAHNEMAN; TVERSKY, 1979), questionam pressupostos da HME, principalmente a racionalidade dos agentes, e conforme Almeida (2011), buscam incorporar aspectos psicológicos ao processo de avaliação e precificação de ativos financeiros, e investigar a influência de eventos e notícias em diferentes mercados (TETLOCK, 2010; SINGER; DREHER; LASER, 2012; ZHANG; SKIENA, 2010). Estes trabalhos indicam que além de dados fundamentais, os eventos, fatos e notícias geradas por eles, são relevantes para o estudo de variação do preço em mercados.

Há ainda, o fenômeno denominado Efeito Manada (CONT; BOUCHAUD, 2000) que diverge da HME, pois é uma ação correlacionada de investidores sem racionalidade provocando variações anormais nos preços dos ativos e influência em retornos esperados. É um fenômeno frequente no mercado

financeiro, investigado em diferentes níveis (SCHARFSTEIN; STEIN, 1990; NAKAGAWA; UCHIDA, 2011).

A construção de modelos de predição e estratégias operacionais é um desafio para economistas e investidores devido à complexidade do fenômeno, que não é explicado completamente com a utilização de análises tradicionais com dados macroeconômicos, microeconômicos e financeiros das empresas, mas envolve também o comportamento irracional dos agentes de mercado. Assim, o tema é objeto de estudos em Finanças Comportamentais, em que pesquisadores investigam se os agentes seguem o comportamento de massa, imitando irracionalmente o comportamento de outros quando negociam ativos financeiros.

Segundo Damodaran (2006), os testes de eficiência de mercado devem buscar descobrir o quanto o mercado é eficiente, e não simplesmente se ele é ou não eficiente. Portanto, é uma questão em discussão.

Kutchukian (2010) apresenta uma série de trabalhos que apontam graus de ineficiência do mercado e irracionalidade dos agentes, principalmente pela ação de amadores que exibem comportamento de manada. Seu trabalho, aplicado a gestores de fundos de ações, confirma que a informação e as expectativas dos investidores não são homogêneas, e que os investidores são influenciáveis pelas decisões de outros investidores. Este comportamento é observado tanto para investidores profissionais como para amadores que atuam diretamente no mercado de ações.

Cont e Bouchaud (2000) concluem que investidores amadores exibem comportamento de manada, e este comportamento influencia o preço das ações.

Estes trabalhos indicam que os investidores não tomam decisões baseados apenas em utilidade esperada e expectativas de fluxos de caixa futuros, mas também com base nas decisões de outros investidores. Eventos, e as notícias geradas por eles, têm impacto na variação do preço das ações.

Na cafeicultura (MELO; MATTOS, 2012), inferem que as informações são assimétricas no mercado de café e que notícias negativas sobre esses fatores afetam diretamente a produção e contribuem, de forma expressiva, na volatilidade da base do produto. Além disso, Nunes, Saes e Brando (2004) apontam que surtos de alta volatilidade têm duração limitada, e choques na base, positivos ou negativos, demoram um tempo considerável para se dissiparem (MELO; MATTOS, 2012).

Paralelo à discussão sobre eficiência de mercados e com a necessidade de tomar decisões em mercados financeiros, existem escolas de análise com o objetivo de compreender como e porque os movimentos de alta ou baixa ocorreram e qual a probabilidade de se repetirem e em qual intensidade, entre elas: a Análise Fundamentalista e a Análise Técnica (MARQUES; MELLO; MARTINES FILHO, 2006), (ARONSON, 2011).

No mercado de *commodities*, a Análise Fundamentalista tem como base a teoria de que o preço de uma *commodity* representa o ponto de equilíbrio entre sua oferta e demanda. O analista deve identificar quais fatores influenciam oferta e demanda com possíveis impactos no preço.

Pela análise técnica, o analista estuda padrões recorrentes nos preços passados com a intenção de prever o movimento futuro dos preços. A análise inclui vários padrões, sinais, indicadores e estratégias operacionais (ARONSON, 2011).

Mesmo sob críticas pela perspectiva da HME e apesar de fragilidades e limitações, como a ambiguidade dos padrões na análise técnica e a complexidade de analisar em tempo todas as variáveis da análise fundamentalista, ambas são utilizadas no mercado para identificar pontos de compra e venda como parte de estratégias operacionais que adotam critérios para gerenciamento de risco. E, no cenário de incerteza do mercado e complexidade

para lidar com várias variáveis, as abordagens para tomada de decisão exigem a capacidade de extrair e interpretar informação de grandes volumes de dados.

3.3 Dados estruturados, não estruturados e conhecimento

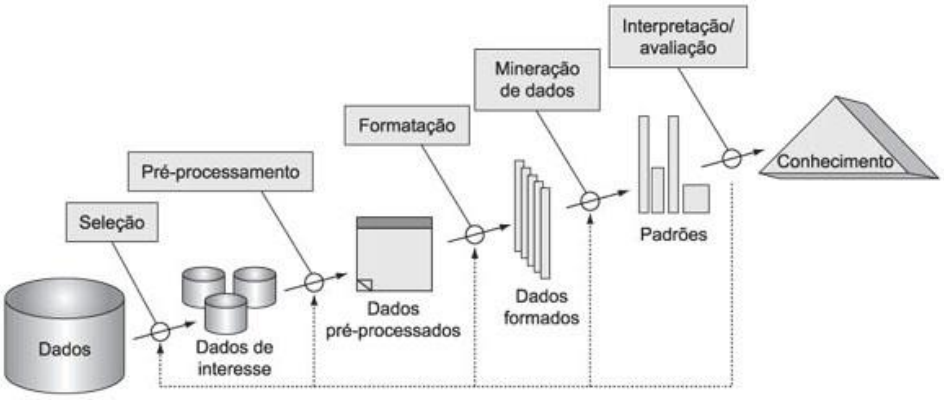
A informação é um conceito chave neste trabalho e as tarefas de monitoramento – coleta, tratamento e análise, e transformação em conhecimento. No ambiente digital, a disponibilidade de informação cresce exponencialmente e lidar com este volume é um desafio para as organizações. Não basta ter acesso à informação é necessário filtrar o que é útil.

Assim, definição e uso do conhecimento formam uma das bases para o trabalho. Para definição, duas premissas apresentadas por (NONAKA, 1994), influenciam a pesquisa em organização baseada no conhecimento: conhecimento tácito e explicito. Eles podem ser conceitualmente discriminados ao longo de um contínuo e a conversão de conhecimento explica, teórica e empiricamente, a interação entre as formas de conhecimento tácito e explícito.

Quanto ao uso do conhecimento, (NONAKA; VON KROGH, 2009) propõem a pergunta: Qual o resultado da conversão de conhecimento? Como uma das respostas, destacam que o resultado é uma capacidade de ação. Afirmam que novo conhecimento individual adquirido promove melhores ou novas definições ou problemas e soluções e melhor desempenho efetivo em tarefas. E o compartilhamento deste conhecimento permite tomadores de decisão nas organizações projetar cenários com base em novas percepções – *insights*.

A concretização de sistemas de informação nas organizações reforçou a necessidade de descoberta de conhecimento a partir dos dados – KDD (*Knowledge Discovery in Databases*) é um processo de extração, em bases de dados, de relações não observadas por especialistas e a validação de conhecimento extraído. Uma de suas etapas é mineração de dados (*Datamining*), Figura 7, onde a matéria prima são os dados.

Figura 7 - Transformação de dados em conhecimento.



Fonte: (FAYYAD et al., 1996).

Dados podem ser vistos como: estruturados e não estruturados. Dados estruturados são aqueles organizados em uma estrutura modelada previamente para atender requisitos de um negócio. A organização facilita a recuperação dos dados. Um exemplo são bancos de dados organizados em sistemas de gerenciamento, acessados por linguagem padrão – *Structure Query Language* (SQL), enquanto dados não estruturados não possuem esta organização e forma padrão de acesso, por exemplo, dados de sensores, áudio e texto.

A capacidade de capturar e analisar dados não estruturados é cada vez mais exigida, principalmente com o advento de Big Data – crescimento exponencial de grandes e complexos volumes de dados. Computação, Processamento de Linguagem Natural e Mineração Textual são campos férteis para pesquisas e aplicações. Para as organizações, o desafio é integrar dados não estruturados para tomada de decisões estratégicas.

Neste contexto, o foco desta pesquisa está nos dados não estruturados, uma vez que o conhecimento pretendido terá como ponto de partida notícias da *web* – texto em linguagem natural.

3.4 Tomada de Decisão, *Business Intelligence* e Inteligência Competitiva

De acordo com Melé (2010), Tomada de Decisão é um processo no qual as pessoas definem um ou mais objetivos para resolver um problema pré-definido. Para atingir estes objetivos, um conjunto de ações é gerado e uma ação é selecionada a partir de avalições e comparações, segundo critérios ou padrões, executada e avaliada.

A escolha de um plano de ação em diferentes cenários e variáveis é um fator determinante de desempenho nas organizações, o que impulsiona pesquisas sobre Tomada de Decisões. Estas pesquisas procuram entender como indivíduos tomam decisões em condições de incerteza, portanto é relevante para vários campos além da Administração e conduzida sob diferentes perspectivas (MELÉ, 2010; OLAREWAJU, 2012; SHEPHERD; WILLIAMS; PATZELT, 2014; BAKHT; EL-DIRABY, 2015).

Sob a perspectiva de tomadores de decisão, ferramentas de decisão e técnicas de seleção analisados nas dimensões epistemológica, teórica e de aplicação Bakht e El-Diraby (2015) apontam que o perfil dos tomadores de decisão evoluiu de indivíduos para uma hierarquia e recentemente para uma rede de tomadores de decisões, e conforme Dai (2013), em um ambiente competitivo, processos reais de tomada de decisões não são realizados de acordo com um procedimento específico pré-definido, mas sim contam com tarefas em vários estágios simultaneamente. É um processo interativo com definição de problema, avaliação do problema, escolha de um plano, feedback e correção.

Os autores concordam que o progresso da Tecnologia da Informação contribui para o crescimento da tomada de decisões estratégicas. E como ressalta Olarewaju (2012) à medida que os ambientes se tornam mais competitivos, tomada de decisões estratégicas competitivas também se tornam mais relevantes. E para suportar os diferentes níveis estratégicos de tomada de decisões, as

empresas precisam explorar informações internas e externas, para promover inteligência para tomada de decisões estratégicas.

Inteligência requer alguma forma de análise cujo propósito é extrair algum significado de dados e informações (BOSE, 2008). À medida que aumentou o volume de dados e informações sobre empresa e ambiente externo, aumentou a necessidade de promover inteligência, o que impulsionou o surgimento e avanço das áreas de *Business Intelligence* (BI) e Inteligência Competitiva (IC).

Desde que o conceito de BI foi introduzido pelo *Gartner Group of Howard Dresner*, em 1989 (SHI; PENG; XU, 2012), seu mercado cresceu mesmo em períodos de recessão (TRUJILLO; MATÉ, 2012), pois BI auxilia a alocação racional de recursos e fluxo de informação a tempo para tomada de decisões. É uma solução que traz vantagem competitiva aos gestores pelo uso de tecnologias que permitem extrair conhecimento útil sobre áreas críticas do negócio a partir dos dados da organização.

A pesquisa acadêmica apresenta BI como um termo geral – guarda-chuva (LIM; CHEN; CHEN, 2013; SELENE XIA; GONG, 2014) sob o qual estão os conceitos e métodos para promover a tomada de decisão usando sistema de apoio baseado em fatos e arquiteturas subjacentes, ferramentas, banco de dados, aplicações e metodologias. Bose (2009) apresenta duas visões: a visão gerencial – focada em entregar a informação certa para as pessoas certas no momento certo para tomarem decisões que melhorem o desempenho das empresas, e a visão técnica – inclui processo, aplicações e tecnologias para coletar, organizar, armazenar e prover acesso a dados para ajudar nas melhores decisões de negócios.

Como uma abordagem focada em dados, BI depende de avanços de *business analytics* (BA) – conjunto de tecnologias que tratam de coleta, extração e análise de dados. Disto, somado ao advento de *Big Data*, emerge o termo

Business Intelligence and Analytics (BIA) como importante área de pesquisa para organizações contemporâneas (CHEN; CHIANG; STOREY, 2012; LIM, CHEN; CHEN, 2013; WIXOM et al., 2014).

Na visão geral, BIA se refere a tecnologias, sistemas de práticas e aplicações que analisam os dados críticos para auxiliar uma empresa a entender melhor seu negócio e mercado. Um estudo abrangente de BIA é encontrado em Chen, Chiang e Storey (2012) que apresenta a evolução em termos de características chaves e capacidades, aplicações, desafios e oportunidades para a pesquisa, com destaque para mineração textual e *web*, assim como em Bose (2009).

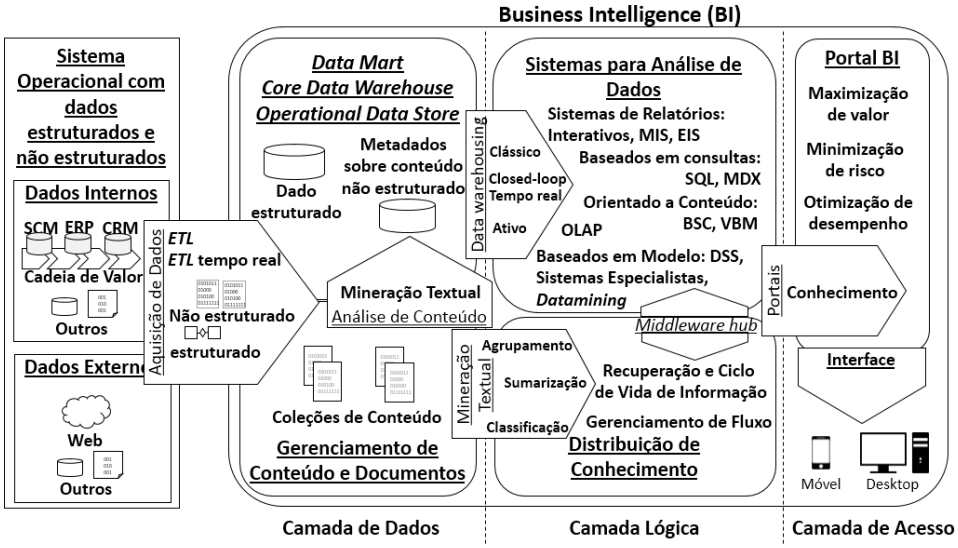
Na visão técnica, as tecnologias incluem *data warehouse*, *data mining*, *online analytical processing* (OLAP) em arquiteturas com camadas que vão desde a aquisição de dados até o gerenciamento de desempenho corporativo. Não há uma arquitetura padrão ou um conjunto de componentes obrigatórios, existem diferentes modelos de BI, variando a estrutura de acordo com as diferentes aplicações e organizações. É notório que com o crescimento da *web* e disponibilidade de dados dentro das organizações, as arquiteturas de BI clássicas incorporaram tecnologias para recuperar e analisar dados não estruturados.

Como referência neste trabalho, tem-se o modelo conceitual em três camadas apresentado por Baars e Kemper (2008), adaptado com base em Bose (2009) e Wu (2010) para descrever arquitetura e processo de BI (FIGURA 8).

Na camada de dados, estão os repositórios para os dados estruturados – *Data Warehouses*, *Data Marts* e *ODS* (*Operational Data Stores*) e não estruturados – Sistemas de Gerenciamento de Conteúdo e Documentos, coletados via ferramentas de extração, transformação e carga (*ETL – Extract Transform Load*) de sistemas internos como *ERP – Enterprise Resource Planning* e *SCM – Supply Chain Management* e de externos – documentos, planilhas e conteúdo da *web*. Os dados estruturados são transformados em

informação para a camada lógica por *data warehousing* e os dados não estruturados por técnicas de análise de conteúdo apoiadas por computador e Mineração Textual.

Figura 8 - Modelo de *Business Intelligence.*



Fonte: adaptado de Baars e Kemper (2008), Bose (2009) e Wu (2010).

A camada lógica tem funcionalidades para analisar a informação e distribuir conhecimento relevante para a camada de acesso, usualmente, um portal com interface para o usuário interagir com o sistema segundo propósito da aplicação. Existem modelos de BI em vários domínios: e-learning, bancos e indústria; para atender diferentes requisitos de negócios como: apoiar a análise de risco, melhorar a confiabilidade, aprimorar a relação com clientes entre outros descritos por Martin, Lakshmi e Venkatesan (2012).

Além de estrutura tecnológica de BI, é necessário conhecimento sobre o negócio aliado ao domínio de estatística e modelos matemáticos que

possibilitem a busca de padrões nos dados para apoiar análises preditivas (extrapolação) e prescritivas (simulações e cenários).

Uma aplicação no domínio de BI é Inteligência Competitiva (BAARS; KEMPER, 2008). IC é usada nas empresas para vários tipos de decisões como desenvolvimento de produtos, entrada em mercados, pesquisa e desenvolvimento e desenvolvimento corporativo. IC promove vantagem competitiva (BRODY, 2008) e desempenho das empresas (SHI, 2011), portanto, assim como BI, cresceu no mercado e academia, como aponta Calof, Richards e Smith (2015). Não é raro IC aparecer como sinônimo de BI. Entretanto, literatura e mercado distinguem os termos, que evoluíram independentes.

Pela perspectiva de dados, enquanto BI é inteligência sobre a própria empresa, IC é inteligência sobre o ambiente externo e competidores (Bose, 2008). Pelo ângulo prático, enquanto BI foca em modelagem e projeto de banco de dados e ferramentas de TI, IC exige pensamento estratégico e metodologias científicas de análise. Em comum têm o objetivo geral de prover informação útil para tomada de decisão estratégica – desafio para as organizações diante do crescente volume de dados coletados de fontes internas e externas como a internet, o que reforça a tendência de integrar estruturas de TI para IC (VEDDER et al., 1999).

Do ponto de vista teórico, Inteligência Competitiva é um campo de interseção das áreas de Ciência da Informação, Gestão do Conhecimento, Engenharia de Requisitos e Sistemas de Informação. Evoluiu também da economia, do *marketing*, da teoria militar e do gerenciamento estratégico (MULLER, 2006), portanto existem várias definições na literatura (PELLISSIER; NENZHELELE, 2013a). De acordo com Calof, Richards e Smith (2015), as definições têm foco nos objetivos ou no processo. Quanto aos objetivos, cita a definição da *SCIP – Strategic and Competitive Intelligence Professionals*: “Disciplina necessária para a tomada de decisões éticas com base

na compreensão do ambiente competitivo" e de Du Toit (2013): "IC é uma ferramenta estratégica para facilitar a identificação de oportunidades e ameaças potenciais".

E com foco no processo, Kahaner (1997) afirma que é "um programa sistemático para coletar e analisar informações sobre atividades dos concorrentes e tendências gerais de negócios para atingir os objetivos da empresa".

Processo e objetivo são encontrados na definição de Brody (2008): o processo pelo qual empresas reúnem informação sobre competidores e ambiente competitivo, e a aplicam em seu planejamento e tomada de decisões para promover o desempenho da empresa.

Características comuns na evolução das definições são: processo sistemático, ética, transformação de dados em conhecimento estratégico (TARAPANOFF; GREGOLIN, 2002), capacidade de utilizar informações públicas (GOMES; BRAGA, 2004), e como ressaltado por Rapp, Agnihotri e Baker (2011), a maioria das definições de IC tem foco em concorrentes ou no ambiente competitivo e resulta em conhecimento que pode ser usado para ganhar vantagem competitiva e antecipar ações. Outra característica imprescindível, reforçada com *Big Data*, é a capacidade para lidar com dados não estruturados.

Para IC é necessário o desenvolvimento de um sistema formalmente organizado para apoio a tomada de decisões (ANICA-POPA; CUCUI, 2009) e um processo apropriado (DU TOIT; MULLER, 2004). Para Saayman et al. (2008), um processo tem as etapas de planejamento, coleta, análise e comunicação de inteligência, em um ciclo contínuo no qual o dado bruto externo é adquirido, transmitido, avaliado, analisado e disponibilizado para tomada de decisão.

Como apontado por Pellissier e Nenzhelele (2013b) existem vários modelos de processos para IC que variam quanto ao número, nome e interação

de fases. Buscando uma definição unificada a partir de uma revisão da literatura os autores sugerem o modelo ilustrado na Figura 9.

Figura 9 - Processo de IC proposto por Pellissier e Nenzhelele (2013b).

Os autores relatam que a maioria dos trabalhos acadêmicos apresentam o processo de IC como um ciclo de fases inter-relacionadas. A saída de uma fase é entrada para a próxima fase. As fases são influenciadas por tomadores de decisões, processo e estrutura, consciência e cultura organizacional e *feedback*.

O processo de IC é apresentado como um ciclo, cada fase é suscintamente descrita a seguir:

1. Planejamento e Direção: Nesta fase, definem-se os requisitos de inteligência claros e não ambíguos dos tomadores de decisão – Tópicos

Fundamentais de Inteligência (HERRING, 1999), que serão transformados em requisitos de informação que pode já existir ou ser coletada em passos claramente delineados.

2. Coleta de Informações: O foco nesta fase é coletar informação pública disponível relevante para os requisitos de inteligência definidos, de fontes primárias (funcionários, clientes, fornecedores, etc.) e secundárias (revistas, TV, relatórios profissionais, análises), via buscas na *web*, formulários, entrevistas, observação e digitalização de mídia.

3. Triagem, captura e armazenamento de informações: A informação coletada é organizada em mecanismos de armazenamento projetados e implementados nesta fase

4. Análise de Informações: Nesta etapa, a informação processada é interpretada e analisada para produzir inteligência. Alguns métodos utilizados são PESTEL Analysis (*political or legal, ecnonomical, socio-cultural and technological*), análise de cenário, Forças Competitivas de Porter, *SWOT Analysis* (*strengths, weaknesses, opportunities and threats*) (VIVIERS; SAAYMAN; MULLER, 2005), *Event and Timeline Analysis* (*ETA*), fatores críticos de sucesso (FLEISHER; BENSOUSSAN, 2014), *Balanced Scorecards*, *Benchmarking*, *Datamining* e *Data warehousing* (TARAPANOFF; GREGOLIN, 2002).

5. Difusão de Conhecimento: O objetivo nesta fase é disseminar inteligência para os tomadores de decisão em um formato que seja facilmente compreendido – relatório, *dashboard* ou *intranet.*

Exemplos de características dos competidores analisadas para IC: capacidade de serviços, alianças ou *joint ventures*, planos futuros e estratégias para mercados específicos, linhas de produtos, razões para mudanças

corporativas ou unidades estratégicas de negócio (BRITT, 2006 apud BOSE, 2008).

3.5 Ferramentas e métodos de inteligência competitiva

Existem métodos clássicos de análise para IC que permitem colocar os dados coletados em um contexto útil para tomada de decisão estratégica, conforme afirma Dai (2013).

O avanço da tecnologia usada em BI, como Mineração *Web* e Textual, contribui para a automatização de métodos utilizados para IC, promovendo precisão, escala e desempenho nas análises. Entre os métodos estudados serão descritos dois utilizados no experimento desta pesquisa: *Event and Timeline Analysis* (*ETA*) e *SWOT*.

3.5.1 *Event and Timeline Analysis - ETA*

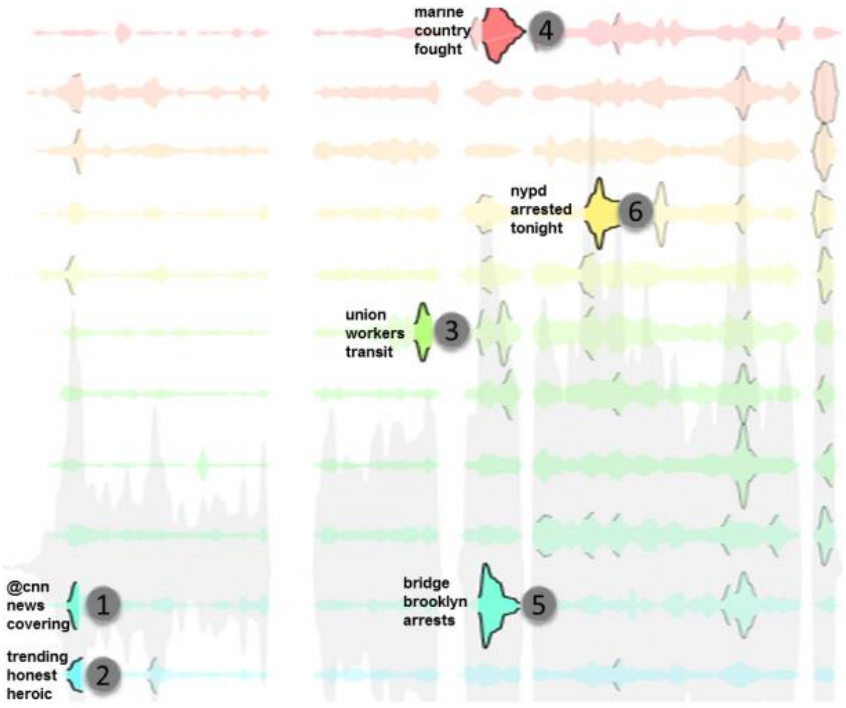
A influência de eventos no desempenho de empresas e efeitos em mercados é amplamente estudada na economia e finanças. Na IC, eventos sobre o ambiente externo, competidores ou comportamento dos atores como consumidores, parceiros e fornecedores, auxiliam a identificar tendências no negócio. O uso de ETA é reforçado para explicar e predizer o desenvolvimento de indústrias e corporações em seu ambiente dinâmico e complexo e pela necessidade de lidar com sobrecarga de informação que diariamente é digitalizada e capturada pela tecnologia de recuperação.

ETA é um grupo de técnicas que exibem eventos sequencialmente. Isola eventos externos e destaca as tendências, semelhanças, e aberrações no comportamento de concorrente ou outros atores. Quando feito de forma sistemática, ETA pode descobrir tendências importantes sobre o ambiente competitivo de uma empresa e servir como uma função de alerta precoce, destacando quando um concorrente normal ou outro ator está desviando do curso de seu comportamento (FLEISHER; BENSOUSSAN, 2014).

Sequências específicas de eventos em uma linha de tempo podem sugerir relações entre eventos em um determinado contexto. As técnicas de análise têm a forma de gráficos, tabelas, diagramas e séries temporais com eventos em ordem cronológica que permitem descobrir padrões e tendências sobre os competidores e ambiente pela relação entre os eventos. Com os avanços da Computação, em Mineração *Web* e Textual, existem trabalhos relacionados como: detecção de eventos (Event Detection) (HUANG et al., 2014; DOU et al., 2012) e de tópicos (Topic Detection and Tracking) (ALLAN, 2012; CHEN; LUESUKPRASERT; CHOU, 2007), evolução de eventos (QIU et al., 2008), que promovem a capacidade de análise total ou parcialmente automatizada para grandes volumes de dados.

Dou et al. (2012) destacam que as pessoas conseguem segmentar atividades físicas observadas em eventos, facilmente e simultaneamente em múltiplas escalas de tempo como apontam Kurby e Zacks (2008), entretanto, há pouca evidência que indique a mesma habilidade aplicada a fluxos contínuos abstratos como tópicos derivados de texto. Um exemplo de aplicação é apresentado pelos autores por meio de um sistema que permite identificar automaticamente da *web* eventos significativos cronologicamente estudados durante a ocupação de *Wall Street* em 2011 (FIGURA 10).

Figura 10 - Principais picos de eventos durante a ocupação de Wall Street.



Fonte: (DOU et al., 2012).

Há um processo para a aplicação das técnicas que começa com a definição de uma linha de tempo e eventos, coleta de dados, organização cronológica dos dados, visualização e conclusões (FLEISHER; BENSOUSSAN, 2014). A evolução do processo e ferramentas contribui para a IC ao passo que revela informações que não seriam facilmente capturadas por analistas diante de um fluxo contínuo de dados textuais. A contribuição dessas ferramentas de análise foi explorada neste trabalho para apoiar a aquisição de IC para tomada de decisões na cafeicultura.

3.5.2 *SWOT Analysis*

Um dos requisitos para tomada de decisões, planejamento e construção de estratégias em uma organização é a capacidade de reconhecer suas competências, pela identificação de fatores internos e externos que auxiliam ou impedem o alcance de objetivos. No âmbito da IC, a análise SWOT é uma das ferramentas utilizadas no mercado e pesquisada pela comunidade acadêmica com este propósito (BOSE, 2008). Considerando sua definição geral: é um método básico para analisar e posicionar os recursos de uma organização e ambiente em quatro regiões: forças (Strengths), fraquezas (Weakness), oportunidades, (Opportunities) e ameaças (Threats) (SAMEJIMA et al., 2006). No caso, forças e fraquezas são fatores internos, considerados controláveis. Enquanto oportunidades e ameaças são fatores externos, considerados incontroláveis (HILL; WESTBROOK, 1997).

Um dos diagramas mais conhecidos da análise é a Matriz SWOT (Dai; KAKKONEN; SUTINEN, 2011a; PHADERMROD; CROWDER; WILLS, 2014; CHOUDER; CHALAL, 2014), apresentada na Tabela 1. A partir da construção da matriz, é possível encontrar sugestões sobre qual estratégia escolher (SO, WO, ST ou WT) combinando diferentes fatores com diferentes objetivos.

Tabela 1 - Matriz SWOT.

Strategy	***Strenghts*** ***S1*** ***S2***	***Weaknesses*** ***W1*** ***W2***
Opportunities ***O1*** ***O2***	*SO strategy* *S1O1, S1O2* *S2O1, S2O2*	*WO strategy* *W1O1, W1O2* *W2O1, W2O2*
Threats ***T1*** ***T2***	*ST strategy* *S1T1, S1T2* *S2T1, S2T2*	*WT strategy* *W1T1, W1T2* *W2T1, W2T2*

Fonte: (DAI et al., 2011a).

Para aumentar o poder de análise, as pesquisas propõem adicionar técnicas de tomada de decisão multicritério ao modelo básico, como Analytic Hierarchy Process (AHP) (GÖRENER; TOKER; ULUÇAY, 2012). Assim como em ETA, as pesquisas utilizam as tecnologias de Mineração Textual para identificação de fatores SWOT. No sistema proposto por Dai (2013) a Mineração Textual é utilizada para extrair fatores SWOT de dados não estruturados internos – *e-mails* e relatórios da empresa, e externos – respostas de clientes e comunicados à imprensa de competidores. Pai et al. (2013) desenvolveram um mecanismo que utiliza Análise de Sentimento e Ontologia para classificar as avaliações *online* e interpretá-las em fatores SWOT.

Como crítica ao método, Phadermrod, Crowder e Wills (2014) ressalta que SWOT em alguns casos pode levar a decisões erradas por ser baseada em análise quantitativa em que os fatores carregam visão subjetiva pelo julgamento dos gestores e não são classificadas pela importância para o desempenho da organização. Isto mostra a necessidade de definição prévia dos objetivos relacionados aos fatores SWOT para estudo em IC. No caso deste trabalho, o contexto da cafeicultura.

Os conceitos Estratégia, BI e IC, e as ferramentas de análise apresentados são realizados, apoiados e potencializados com o uso de tecnologias em constante evolução, principalmente pela área de computação e suas ramificações em Tecnologia da Informação e Sistemas de Informação e interseção com áreas como Administração, Economia e Ciências Sociais, especialmente em aplicações com dados não estruturados. As principais tecnologias utilizadas neste trabalho, citadas juntamente com os conceitos apresentados, serão descritas a seguir: Mineração de Dados, Mineração Textual e Análise de Sentimento.

3.6 Mineração de dados (*Data Mining*)

A humanidade vivencia uma explosão de dados impulsionada pela redução do custo e aumento da capacidade tecnológica de coleta, armazenamento e compartilhamento, aliada ao número crescente de usuários e formas de produzir dados em diferentes formatos. Esse volume de dados cresce à ordem de *zettabytes* (10^{21} bytes) e constitui a base do fenômeno denominado *Big Data* (MANYIKA et al., 2011; ZIKOPOULOS; EATON, 2011).

Um problema neste cenário é a quantidade versus qualidade. Conforme Bramer (2013) o mundo está se tornando "rico em dados, mas pobre em conhecimento" uma vez que grande parte destes dados é apenas armazenada e não analisada. Imerso no volume pode estar o conhecimento para novas descobertas e avanços. Para as organizações, especialmente as baseadas em conhecimento, saber lidar com este problema é fator de sucesso ou fracasso. O uso de Mineração de Dados é uma das soluções que permite descobrir padrões nos dados, dificilmente encontrados com métodos tradicionais de consulta.

A definição geral de Mineração de Dados para a administração é o uso da tecnologia da informação para descobrir regras, identificar fatores e tendências, descobrir padrões e relacionamentos ocultos em grandes bancos de dados para auxiliar a tomada de decisões sobre estratégia e vantagens competitivas. É a parte central da extração de conhecimento (*Knowledge Discovery*).

Apesar de ser considerado sinônimo de Mineração de Dados, para alguns autores, o termo extração de conhecimento é definido por Piateski e Frawley (1991) como a extração não trivial de informação implícita, previamente desconhecida e potencialmente útil, a partir de dados.

Em Kantardzic (2011), a Mineração de Dados é dividida, quanto aos objetivos, em preditiva e descritiva. Enquanto preditiva, o objetivo é produzir um modelo descrito por um conjunto de dados que pode ser usado para

classificação, predição, estimação e outras tarefas similares. A descritiva produz informação nova, não trivial, baseada no conjunto de dados disponível com o objetivo de obter entendimento pela descoberta de padrões e relacionamentos em grandes conjuntos de dados.

O conjunto de dados pode ser rotulado ou não rotulado. No primeiro caso, existe determinado atributo com valores atribuídos para diferentes instâncias do conjunto e a tarefa é predizer o valor deste atributo para novas instâncias não rotuladas previamente, com base nas instâncias existentes. No segundo caso, não há valor previamente determinado para atributo de instâncias do conjunto de dados, e o objetivo é extrair o máximo de informação das instâncias disponíveis.

Cada tipo de dados terá uma técnica adequada, dentre elas: classificação, previsão numérica, associação e agrupamento (*clustering*). A proposta neste trabalho é a classificação, pertinente, pois o processo de análise do BIC constrói uma base de dados rotulada adequada à técnica.

Classificação é a tarefa de dividir objetos em categorias mutuamente exclusivas, conhecidas como classes. Um exemplo prático é a classificação realizada por gerenciador de *e-mails* que identifica uma mensagem como legítima ou *spam*. Existem diferentes algoritmos para a classificação, parte investigada neste trabalho.

Segundo Bose (2008), a tendência dos avanços recentes em *datamining* é a Mineração Textual, que aplica as mesmas funções para informação textual resultando em técnicas para análise de texto.

A base de dados para este trabalho é constituída de textos – informação textual presente em notícias. Para extrair conhecimento deste tipo de dado não estruturado e com alto grau de subjetividade, a Mineração de Dados conta com os avanços na especialidade Mineração Textual ou *Textmining*. Embora este tema tenha sido reforçado por *Big Data*, o texto como dado para inferências,

manual ou com o uso de computadores, tem sido estudado há anos em outras áreas – política, por exemplo (GRIMMER; STEWART, 2013; LUCAS et al., 2015). A Análise de Conteúdo é apresentada na próxima seção como técnica para tratamento de texto.

3.6.1 Tratamento Sistemático de Texto

A análise de conteúdo é uma abordagem qualitativa para tratamento sistemático de texto, é uma técnica das ciências da comunicação desenvolvida nas primeiras décadas do século XX nos EUA, para analisar meios de comunicação em massa (jornais, rádio) (MAYRING, 2000; BARDIN, 2006). A análise evoluiu do campo quantitativo de frequência como contagem, atribuição de pesos e relacionamento entre os elementos do texto, para uma análise de conteúdo considerando contexto e estruturas de sentido latentes.

A análise de conteúdo consiste em um conjunto de técnicas de análise das comunicações, que utiliza procedimentos sistemáticos e objetivos de descrição do conteúdo das mensagens. A intenção da análise de conteúdo é a inferência de conhecimentos relativos às condições de produção (ou eventualmente, de recepção), inferência esta que recorre a indicadores quantitativos ou não (BARDIN, 2006).

A ideia central nas técnicas é um sistema de categorias desenvolvido a partir do material e da teoria, o qual determina os aspectos que devem ser filtrados do material. Uma forma é a Estruturação, definida por Mayring (2000) com o objetivo de estabelecer um recorte do material na base de critérios pré-estabelecidos. Bardin (2006) divide a análise de conteúdo em três fases em torno dos polos cronológicos: pré-análise, exploração do material, tratamento dos resultados, inferência e interpretação.

A fase pré-análise tem como objetivo organizar e sistematizar as ideias iniciais para definir um programa preciso de análise. Para isso, há as atividades,

não em ordem cronológica: escolha dos documentos, formulação de hipóteses e objetivos e elaboração de indicadores para a interpretação final. A partir da formulação de hipóteses e objetivos determina-se quais os elementos do texto devem ser considerados e como recortar o texto em elementos completos por meio de unidades de registro e contexto.

A unidade de registro tem natureza e dimensões variáveis: tema (nível semântico), palavra, palavras ou frase (nível linguístico). Representa a unidade de significação codificada e corresponde ao segmento de conteúdo considerado unidade básica para categorização e contagem de frequência. E a unidade de contexto é aquela de dimensão superior, que serve de unidade de compreensão para codificar a unidade de registro, exemplos: frase para a palavra, parágrafo para o tema.

Aplicando regras de enumeração, a divisão das componentes das mensagens é analisada em categorias. A categorização é uma classificação de elementos constitutivos de um conjunto por diferenciação e, em seguida, por reagrupamento segundo o gênero (analogia), com os critérios previamente definidos. As categorias são rubricas ou classes, as quais reúnem um grupo de elementos sob um título genérico, agrupamento esse efetuado em razão de características comuns destes elementos (BARDIN, 2006).

A categorização pode ser a partir de categorias pré-definidas, nas quais os elementos são classificados ou resultantes de classificação analógica e progressiva sem um sistema de categorias pré-definido. Estas formas são básicas para a técnica de análise categorial, entretanto, a maioria das técnicas de análise tem o processo de categorização. Existem outras técnicas, entretanto não é objetivo descrevê-las, podem ser vistas em Mayring (2000) e Bardin (2006) que apresenta: análise de avaliação, análise da enunciação, análise proposicional do discurso, análise da expressão e análise das relações.

Para ilustrar, considera-se um corpus simples, formado por apenas quatro notícias sobre o mercado de café, selecionadas por especialistas, como unidade de registro, o radical das palavras: *production (product)*, *export*, *machine*, *farm*, *capsule*, *manufacture (manufactur)* e como unidade de contexto, os parágrafos das notícias – unidades de compreensão para identificar as palavras unidades de registro. A Tabela 2 mostra um exemplo de classificação pela identificação das categorias: Indústria – com notícias com fatores que influenciam a indústria de café, como lançamentos de produtos e parcerias entre empresas e produção – com notícias que devem ser estudadas no escopo da produção de café, como fatores climáticos e pragas que afetam a produção.

Tabela 2 - Análise de Conteúdo.

Título da Notícia	Unidades de Contexto e Registro	Categorias
Italy's Lavazza may raise Green Mountain stake	*"Green Mountain Keurig coffee brewers compete with Nestle's Nespresso machines. Mele said Lavazza and Green Mountain planned to launch a new capsule coffee machine which will make cappuccino, or creamy milk coffee, on the U.S. market by May 2012."*	Indústria
Starbucks, Green Mountain In Talks To Forge Coffee Partnership	*"Starbucks reportedly said last week that it plans to announce a new product for the single-cup coffee market in the near future, and is expected to make its own vending machines or partner with a manufacturer of coffee machines."*	
Colombia coffee production recovers	*"... even though coffee exports returned to normal levels in January after a sharp dip,... show the Colombian coffee industry, which was hit by the heavy rainy season and fungus outbreaks, produced 908,000 60-kilogram bags of coffee in the first month of 2011*	
Uganda: Coffee Sector Can Be Revived	*"The Ugandan coffee industry is facing challenges that range from the coffee wilt disease, declining volumes, climate change and lack of funding, which have contributed to the declining coffee exports. ... Huge chunks of coffee farm land have been lost, thus reducing the land for coffee production.*	Produção

A análise poderia incluir subcategorias para refinar o resultado das inferências de acordo com hipóteses definidas anteriormente. O desmembramento do material em categorias tem como objetivo gerar indicações para o processo de inferência sobre regularidades e propriedades que vão contribuir para o processo de interpretação. Os critérios de escolha e de delimitação das categorias são determinados pelos temas relacionados aos objetos de pesquisa e identificados nos discursos dos sujeitos pesquisados (VALENTIM, 2005).

A Análise de Conteúdo é utilizada em estudos na Administração e Ciências Sociais como metodologia rígida para análise qualitativa. Com o surgimento dos computadores, vieram também os *softwares* que permitiram ganhar escala pela análise de grandes volumes de dados, digitalizados ou produzidos em meio digital. Estes *softwares* foram chamados de CAQDAS (*Computer Assisted Qualitative Data Analysis Software*) entre eles: *General Inquirer*, MAXQDA, ATLAS.ti, NVivo e Sphinx.

Entretanto, estas ferramentas foram introduzidas gradativamente para analisar dados qualitativos, diante do risco de adotar uma epistemologia positivista reducionista pelo uso de métodos dos CAQDAS. Muitos pacotes se restringiam a replicar o trabalho manual de codificação e marcação formal (KUCKARTZ, 2004 apud WIEDEMANN, 2013), dividindo as abordagens quantitativa e qualitativa.

Críticos apontam que a quantificação reduz a precisão da análise quando negligencia a exploração qualitativa e o julgamento quanto ao viés que pode surgir pela definição de categorias o que coloca em xeque a capacidade dos *softwares* de incluir contexto. Contexto é essencial na perspectiva qualitativa, pois forma a base para significado e pela perspectiva da computação linguística, é a fonte decisiva para superar a simples contagem de caracteres em modelos mais complexos da linguagem e cognição humana (WIEDEMANN, 2013).

Os paradigmas devem ser integrados para a análise de texto com métodos combinados que reúnam quantificação e modelos de conhecimento que permitam extrapolar a possível influência de um *software* específico (WIEDEMANN, 2013). Para a análise de textos, as aplicações CATA – *Computer-Assisted Text Analysis* devem distinguir o processamento do gerenciamento de dados, agregando à análise de conteúdo aspectos quantitativos como frequência, coocorrência e distância de palavras, e qualitativos como contexto e significado.

Na evolução em busca por contexto e significado, surgiram várias ferramentas computacionais. Wiedemann (2013) apresenta quatro tipologias de CATA baseadas na capacidade computacional de cada época e noção de contexto:

a) CAQDAS (*Computer Assisted Qualitative Data Analysis Software*) – *softwares* independentes de métodos que fornecem ferramentas para processos de codificação manuais admitindo contextos linguísticos e situacionais;
b) CCA (*Computational Content Analysis*) – orientados pela hipótese, produzindo anotações automáticas pela observação de ocorrência de termos, ignorando contexto;
c) Baseados em lexicometria, direcionados a dados e métodos linguísticos permitindo exploração indutiva de padrões de linguagem medindo contextos evidentes de símbolos linguísticos. Utiliza dicionários léxicos;
d) Mineração Textual (*Textmining*) – abordagens que esforçam para extração de significado por meio de aplicação de modelos estatísticos complexos calculando contextos latentes de símbolos de linguísticos.

A Mineração Textual não considera um único conjunto de unidades léxicas e contribui para a integração de análise quantitativa e qualitativa de texto, pode ser aplicada a um conjunto dinâmico de documentos de entrada em fluxo contínuo, permitindo a inclusão de novos dados qualitativos, além disso não se restringe à ocorrência e frequência de palavras, mas cria um modelo estatístico que contribui para a inclusão de informação contextual, importante na perspectiva qualitativa para configurar significado (WIEDEMANN, 2013).

O conceito de Mineração Textual e aplicação como apoio em CATA é apresentado na próxima seção.

3.6.2 Mineração textual (*Text mining*)

Mineração textual é o processo de obter informação a partir de texto em linguagem natural. 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 semiestruturados. Estas técnicas incluem recuperação de informação, análise léxica, reconhecimento de padrões, análise preditiva e processamento de linguagem natural. Mineração Textual utiliza métodos para transformar texto em um dado que pode ser usado em análises preditivas. As tarefas incluem algoritmos para categorização de textos, identificação de entidades, *text clustering* e sumarização.

A categorização de texto para organização de documentos tem como objetivo classificar documentos de textos de acordo com categorias pré-definidas (JOACHIMS, 1998). Sebastiani (2002) descreve a tarefa como uma função:

$$\phi : DxC \rightarrow \{T, F\} \quad (1)$$

Em que $D = \{d_1, d_2, ... d_{|D|}\}$ é o conjunto que representa o domínio de documentos e $C = \{C_1, C_2, ... C_{|c|}\}$ é o conjunto pré-definido de categorias. O valor *T*, atribuído a $<d_j,c_i>$, indica uma decisão de classificar d_j como c_i, e *F* indica que d_j não é classificado como c_i. O classificador é a função que descreve como documentos devem ser categorizados.

A tarefa é considerada supervisionada quando há informação externa sobre a correta classificação dos documentos, como exemplo, um conjunto de dados previamente rotulados com as categorias. É considerada não supervisionada quando não existe referência à informação externa sobre a classificação – dados não rotulados. E semi-supervisionada quando partes dos documentos são rotulados por mecanismos externos.

Estudos mostram que as técnicas existentes alcançam eficiência e precisão satisfatórias quando treinadas em grandes conjuntos de dados (MANNING; RAGHAVAN; SCHÜTZE, 2008; SEBASTIANI, 2002; LEWIS et al., 2004). O problema, conforme Mccallum et al. (1999), é que em aplicações reais, os recursos humanos necessários para a produção manual da base de treinamento dificultam ou inviabilizam o processo.

No exemplo da seção 3.6.1, Tabela 2, utilizando uma ferramenta de Mineração Textual, a soma de ocorrências de cada palavra em uma notícia é representada em uma lista ordenada. Consequentemente o conjunto de notícias é representado por uma matriz matemática – conjunto destas listas, sobre a qual, algoritmos combinam métodos estatísticos com características da linguagem e conhecimento externo sobre o texto (categorias). A Figura 11 mostra um trecho de arquivo no formato exportado pelo *software* WEKA para Mineração Textual. A linha 3 mostra as categorias pré-definidas. Nas linhas de 5 a 22, palavras dos textos, e de 26 a 29 a matriz gerada a partir da ocorrência das palavras – unidades de registro na Análise de Conteúdo.

Figura 11 - Exemplo de representação de texto.

```
@relation 'cim_categories-weka.filters.unsupervised.attribute.StringToWordVe

@attribute cimcategories {industria,producao}
...
@attribute capsule numeric
@attribute coffee numeric
...
@attribute machine numeric
@attribute machines numeric
@attribute manufacturer numeric
@attribute market numeric
...
@attribute product numeric
...
@attribute exports numeric
@attribute farm numeric
...
@attribute fungus numeric
...
@attribute produced numeric
@attribute production numeric
@attribute rainy numeric
...

@data
{1 0.960906,2 0.960906,3 0.960906,4 ... 0.480453,31 0.960906,38 0.199406}
{9 0.960906,10 0.960906,17 0.960906,18 ... 0.960906,36 0.960906,37 0.960906}
{0 producao,38 0.199406,39 0.960906,41 ... 0.960906,79 0.960906,80 0.960906}
{0 producao,38 0.199406,40 0.960906,42 ... 0.960906,81 0.960906,82 0.960906}
```

Os valores nesta matriz numérica indicam a presença de um determinado termo em um documento e a importância ou distribuição dos termos na coleção de documentos, o que configura um modelo (contexto) para a aplicação de algoritmos de classificação.

Algoritmos de aprendizado de máquina aplicados a estes dados deduzem conjunto de regras ou probabilidades estatísticas de características próprias em novos textos, com esta aprendizagem classificam textos desconhecidos.

Dentre os algoritmos amplamente estudados (SEBASTIANI, 2002; YANG; LIU, 1999; KANG; YOO; HAN, 2012) e utilizados como classificadores, destacam-se os pertencentes a três classes: *Naïve Bayes*, Árvore de Decisões e *Support Vector Machines*.

Métodos baseados em Árvore de Decisões decompõem hierarquicamente o conjunto de dados de treinamento nas classes pré-definidas (QUINLAN, 1986) de acordo com as características presentes nos valores de seus atributos e decide em qual partição é mais provável que um determinado texto pertença. Na estrutura básica da árvore de decisão, o nó raiz é a propriedade da amostra, os ramos são os valores da propriedade, os nós internos denotam diferentes atributos e os ramos entre eles os valores possíveis para estes atributos na amostra observada e os nós folhas denotam o valor final.

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 (MCCALLUM; NIGAM, 1998).

Support Vector Machines (CORTES; VAPNIK, 1995; HEARST ET AL., 1998) é uma técnica que tenta particionar o espaço de dados com delimitações lineares ou não lineares entre as diferentes classes e determinar os limites ótimos entre as classes.

Uma aplicação da Mineração Textual é a Análise de Sentimento, tarefa que procura extrair do texto informação subjetiva como emoção, humor e opinião. A análise de sentimento em texto extraído da *web* também é explorada neste trabalho, portanto descrita na próxima seção.

3.6.3 Análise de sentimento

Milhões de páginas na *web* sobre diferentes temas constituem um repositório de dados de escala mundial que cresce exponencialmente em diferentes formas, entre elas: redes sociais, *blogs*, *microblogs* e fóruns de discussão. Este ambiente é propício à recuperação de dados, que inicialmente se concentra em processamento de informações textuais, mineração e recuperação

de informação fatual, como buscas na *web* e classificação de texto, e tem avançado para mineração de opinião ou análise de sentimento, termos comumente utilizados como sinônimos e considerados como aplicações da Mineração de Dados.

A análise de sentimentos é definida por Liu (2010), como o estudo computacional das opiniões, avaliações e emoções das pessoas, expressas em texto, relativas a entidades, eventos e seus atributos. Esta abordagem é adequada em sistemas onde é necessário saber a opinião de outros para a tomada de decisão. Entretanto, são vários os desafios para a construção de sistemas computacionais para este fim (DEY; HAQUE, 2009; LIU, 2010; PANG; LEE, 2008), como a subjetividade e ambiguidade de textos em linguagem natural, informalidade e ruídos: erros gramaticais, pontuação imprópria, abreviações, gírias e palavras erradas o que aumenta a complexidade de extração da informação.

Pesquisas nesta área exploram buscas sintáticas e semânticas com diferentes níveis de classificação de opiniões, como, identificar a polaridade positiva ou negativa, (DAS; CHEN, 2007; TURNEY, 2002; DAVE; LAWRENCE; PENNOCK, 2003; PANG; LEE, 2008), reconhecer a opinião em classes mais específicas como raiva e aversão e sua intensidade (WILSON et al., 2005; YU; HATZIVASSILOGLOU, 2003; WILSON; WIEBE; HOFFMANN, 2005) e como proposto por Choi et al. (2005), identificar também a fonte da opinião para sistemas que respondem a questões da forma: "O que *X* pensa sobre *Y*?".

Conforme apresenta Montoyo, Martinez-Barco e Balahur (2012), em um levantamento sobre o estado da arte em análise de sentimento, a natureza ambígua e subjetiva do termo sentimento resulta em diferentes pesquisas e aplicações. Os autores dividem as pesquisas em análise de sentimento em quatro categorias: criação de recursos para análise léxica, classificação de textos,

extração de opiniões e aplicações em análise de sentimento, cada qual com desafios específicos.

A classificação de textos de acordo com a polaridade ou orientação de seu conteúdo em positivo, negativo ou neutro tem sido estudada para todo o texto, sentença, frase ou palavra. Duas metodologias, apontadas por Martín-Valdivia et al. (2013), são mais adotadas: Aprendizado de Máquina e Orientação Semântica.

Na metodologia de aprendizado de máquina, o problema é modelado como categorização de texto, em que as classes são positiva e negativa (em alguns casos neutra), e os algoritmos classificadores são treinados, em uma coleção de dados previamente rotulada, para classificar novos dados. Esta é a abordagem supervisionada. Pesquisas com aprendizado de máquina são descritas por Pang e Lee (2008), Liu et al. (2013), Tsytsarau e Palpanas (2012).

A outra abordagem citada considera a orientação semântica, positiva ou negativa, de palavras (TURNEY, 2002), para tal faz uso de recursos léxicos como lista de palavras, características linguísticas como adjetivos e advérbios e dicionários com polaridade (TURNEY, 2002; LIU; HU; CHENG, 2005; DING; LIU, 2007; HU; LIU, 2004; KAMPS et al., 2004).

Existem diferentes recursos léxicos disponíveis contendo informação sobre a implicação emocional de palavras, comumente, referem-se à polaridade das palavras: positiva ou negativa. Dentre os recursos léxicos, os mais estudados estão: General Inquirer (STONE; DUNPHY; SMITH, 1966), MPQA Subjective Lexicon (WILSON; WIEBE; HOFFMANN, 2005), WordNet (FELLBAUM, 1998), SentiWordNet 3.0 (BACCIANELLA; ESULI; SEBASTIAN, 2010), Bing Liu's Opinion Lexicon (HU; LIU, 2004) (LIU; HU; CHENG, 2005), sendo os dois últimos os mais citados (CRUZ et al., 2014).

Abordagens com aprendizado de máquina dependem da disponibilidade de dados rotulados para treinamento, o que pode ser uma tarefa inviável em

alguns casos. Orientação semântica, por outro lado, demanda grande quantidade de recursos léxicos, geralmente dependem da língua e o resultado é comprometido em caso de ausência de palavras do recurso léxico no texto (MARTÍN-VALDIVIA et al., 2013), além de não considerar o contexto.

Apesar destas limitações vários projetos têm alcançado avanços e resultados significativos utilizando estes métodos, como apresentado na próxima seção.

3.6.3.1 Aplicações de análise de sentimento

Como ressaltado em Ren et al. (2013), várias pesquisas em finanças e sistemas de informação têm aplicado esta técnica para análise de material textual de empresas com resultados satisfatórios, por exemplo, com métodos de aprendizado de máquina em relatórios financeiros. São citadas aplicações para identificar fatores de risco (HUANG; LI, 2011), prever ganhos futuros (LI, 2010), prever preços de ações (KRAVET; MUSLU, 2013), impacto na liquidez de mercado (MITRA et al., 2011) e detectar fraudes (HUMPHERYS et al., 2011).

Outras aplicações utilizam análise de sentimento em notícias (LLOYD et al., 2005; BALAHUR ET AL., 2013) com resultados relevantes em diferentes contextos: satisfação de clientes (AGUWA; MONPLAISIR; TURGUT, 2012) e sistema de recomendação (GARCÍA-CUMBRERAS; MONTEJO-RÁEZ; DÍAZ-GALIANO, 2013).

Pang e Lee (2008) destacam aplicações no contexto político em busca de intenções de votos: nos negócios, para capturar satisfação de clientes com produtos: no governo, para delinear a opinião pública sobre um projeto e no mercado financeiro para previsões em operações com ativos. Este último campo de estudo tem relação com o trabalho, uma vez que o café é uma *commodity* negociada em bolsa.

Muitos trabalhos aplicam recuperação da informação na *web*, mineração de opiniões e análise de sentimento ao mercado de ações em busca de correlação entre o movimento dos preços e a expectativa dos investidores em relação aos eventos, fatos e notícias sobre determinada empresa. Bollen, Mao e Pepe (2011) utilizaram um sistema semiautomático de mineração de opiniões para determinar sentimentos específicos como tensão, confusão, vigor e fadiga a partir do microblog *Twitter* e encontraram correlação com os períodos do índice americano Dow Jones.

Das e Chen (2007) desenvolveram um Índice de Sentimento a partir de métodos híbridos para a extração de opiniões sobre ações de fóruns utilizando cinco algoritmos com metodologia diferente de classificação de mensagens: *Naïve Classifier*, *Vector Distance Classifier*, *Description-Based Classifier*, *Adjective-Adverb Phrase Classifier* e *Bayesian Classifier*. O método foi aplicado para 24 ações do setor de tecnologia da bolsa americana que compõem o Índice MSH (*Morgan Stanley High-Tech Index*) em 145.110 mensagens publicadas no período de julho a agosto de 2001. Os resultados mostram uma relação significativa entre o Índice de Sentimento e o Índice MSH normalizados, entretanto, aponta que é difícil inferir um poder preditivo do Índice de Sentimento em movimentações diárias de ações individuais. O resultado do Índice de Sentimento para ações individuais não foi estatisticamente significativo.

Lopes et al. (2008) propõem uma aplicação no mercado brasileiro utilizando *PMI (Pointwise Mutual Information)* para extrair de notícias a opinião expressa sobre empresas listadas na BOVESPA como nível 1 em governança corporativa. Antweiler e Frank (2004) apresentaram forte relação entre volume de mensagens, volatilidade e volume de negócios em ações da bolsa americana. Gerow e Keane (2011) mostram resultados de análise de ocorrências de determinados verbos e substantivos em relatórios financeiros publicados na *web*

pelo New York Times, Financial Times e BBC, indicando sua relevância para rastrear movimentos do índice Dow Jones e para previsão de bolhas.

Wolfram (2010), com técnicas de processamento de linguagem natural, conclui que a extração de dados do *Twitter* promove uma vantagem pequena, mas significativa para a previsão de mercados.

Tetlock, Saar-Tsechansky e Mackassy (2008) apresentam o uso de uma medida quantitativa da linguagem para predizer lucro contábil e retorno sobre ações de empresas. As conclusões são que a fração de palavras negativas nas notícias específicas das empresas prevê baixos rendimentos contábeis, os preços das ações caem brevemente com palavras negativas em informação sobre a empresa e o impacto de palavras negativas é maior quando estas estão em informação sobre fundamentos da empresa.

Ren et al. (2013) explora a aplicação de duas variações de métodos de aprendizado de máquina *Naïve Bayes* para análise de sentimento em dados rotulados extraídos de relatórios textuais sobre duas empresas do mercado americano, Microsoft e Coach. O trabalho aponta as situações em que cada variação apresenta resultados satisfatórios e que quando a quantidade de dados rotulados é limitada, a utilização de dados não rotulados melhora o desempenho da classificação.

Zhang e Skiena (2010) utilizam dados quantitativos gerados de blogs e notícias por sistema de análise textual com métodos e processamento de linguagem natural, em um estudo comparativo para compreender como a frequência, polaridade e subjetividade de notícias sobre uma empresa antecipa ou reflete o volume de negócios com as ações da empresa e retorno financeiro. O trabalho mostra evidência concreta de que os dados na mídia são altamente informativos e desenvolveram uma estratégia de operações no mercado baseada no sentimento extraído desta informação que apresentou desempenho consistente em retornos com baixa volatilidade de 2005 a 2009.

Huang e Li (2011) apresenta a identificação automática de 25 tipos de fatores de risco sobre as empresas do mercado americano em seção específica do relatório padrão 10-K da *Securites and Exchance Commision*. O algoritmo multilabel categorical K-nearest neighbor (ML-CKNN) classificou 74,94% dos fatores de risco e acertou 98,75% dos rótulos. O resultado satisfatório corrobora o uso de técnicas de classificação textual no contexto financeiro.

Dai et al. (2013) propõem um modelo de sistema de suporte à decisão que integra tecnologias de processamento de linguagem natural para mineração de opiniões em dados textuais publicados em *sites* de empresas concorrentes e fontes internas da empresa para facilitar a Inteligência Competitiva pela análise de eventos. Entretanto, tem propósito geral e não apresenta correlação entre dados extraídos da *web* e variáveis específicas.

A maioria são aplicações no mercado americano e têm como foco um índice financeiro ou ações e não *commodities*. Procuram a predição de movimentação de preços sem ênfase explícita em Inteligência Competitiva, com exceção de DAI et al. (2013). Apesar da quantidade de trabalhos com esta abordagem aplicados ao mercado financeiro, são menos frequentes aplicações no mercado de *commodities*, tanto para índices como para uma *commodity* específica como café, soja e milho dentre outros.

3.6.4 Ferramentas para Mineração Textual

A crescente atuação de empresas em processamento de linguagem natural, análise textual e análise de sentimento, impulsionadas principalmente por *Big Data*, reforçam o potencial deste tipo de análise, entre elas: IBM (www.ibm.com/jstart/textanalytics), Appinions (http://www.appinions.com/), Beyond the Arc (http://beyondthearc.com/) e no campo de notícias, Thomson Reuters e Bloomberg. Existem vários *softwares* comercias e acadêmicos para Processamento de Linguagem Natural (*NLP – Natural Languague Processing*),

mineração textual, análise textual e de sentimento, disponíveis como ferramentas, APIs (*Application Programming Interface*), aplicações *web* ou serviços *web*.

Pela abrangência de temas estudados, apresentar todas as ferramentas foge do escopo do trabalho, para tanto a literatura apresenta comparações de *softwares* quanto a características como objetivos, funcionalidades e aplicações – Mineração de Dados (MACEDO E MATOS, 2010), Mineração Textual (YANG et al., 2008), Mineração Textual e Análise de Sentimento para Inteligência Competitiva (CHOUDER; CHALAL, 2014), (DAI, 2013). A ferramenta para Mineração Textual utilizada neste trabalho foi a WEKA, brevemente descrita na Tabela 3 que apresenta outras estudadas no contexto do trabalho.

Tabela 3 - Aplicações e Ferramentas. (Continua)

Aplicação	**Descrição**
Text Map http://www.textmap.com/	Ferramenta de busca que utiliza processamento de linguagem natural para identificar e monitorar pessoas, lugares, e outras entidades em notícias.
Lexalytics Salience Engine http://lexalytics.com/software	Ferramenta para análise textual em várias línguas, com várias técnicas de *NLP* para ser integrada em sistemas.
Semantria http://semantria.com	Ferramenta que oferece análise de sentimento integrada a Excel ou como uma *API web*, utiliza como base Salience Engine
Alchemy API http://www.alchemyapi.com/	*API* distribuída como serviço *web* com funções de *NLP* para análise de sentimento.
Rapid Miner http://rapidminer.com/	Ferramenta de propósito mais geral em análise preditiva inclui métodos para mineração de dados e textual.
IBM SPSS Text Analytics www.ibm.com/jstart/textanalytics	Módulo de análise textual e de sentimento da ferramenta estatística SPSS

Tabela 3 - Aplicações e Ferramentas. (Conclusão)

Aplicação	Descrição
SAS http://www.sas.com/	Conjunto de soluções para Big Data e mineração que permite integração com plataforma de computação distribuída para grandes volumes de dados.
GATE http://gate.ac.uk/	Ferramenta gratuita de código aberto para processamento de texto e marcação semântica com uso de ontologias
NLTK –Natural Language Toolkit http://nltk.org/	É uma biblioteca para programação em Phyton com funções de NLP e possui interface para recursos léxicos como Wordnet.
Open NLP http://opennlp.apache.org/	Biblioteca para programação em Java com funções de NLP baseada em Aprendizado de Máquina.
Stanford Core NLP http://nlp.stanford.edu/software/corenlp.shtml	Biblioteca Java com funcionalidades para NLP em língua Inglesa, inclui análise de sentimento e identificação de entidades
Apache Mahout https://mahout.apache.org/	Biblioteca com algoritmos de aprendizado de máquina que podem ser utilizados para NLP e análise de sentimento.
WEKA http://www.cs.waikato.ac.nz/ml/index.html	Aplicação gratuita que inclui vários algoritmos de mineração de dados e aprendizados de máquina, possui API para algoritmos classificadores.
Opfine http://www.opfine.com/	Site que realiza em tempo real e automaticamente análise de sentimento em notícias sobre o mercado americano, índices, ações de empresas e commodities.
Finsents http://www.finsents.com/	Serviço *web* para monitorar o sentimento em notícias sobre o mercado americano, índices, ações e commodities.
Marketprophit http://www.marketprophit.com/	Serviço *web* para monitorar o sentimento na *web* sobre empresas da bolsa americana. Utiliza NLP em mídia social.

Rapid Miner, IBM SPSS Text Analytics e SAS são ferramentas de propósito geral e não têm uma base de conhecimento específica de análise de sentimento para o mercado de café por exemplo. Semantria e Alchemy são API's que podem ser usadas integradas em planilha eletrônica ou como serviço *web*. Não são gratuitas, utilizam um recurso léxico, mas permitem a configuração de um contexto específico para o usuário.

Opfine, Finsents e Marketprophit são serviços na *web* aplicam análise de sentimento em notícias sobre o mercado financeiro, entretanto, tem foco em ações de empresas. Finsents inclui monitoramento de notícias sobre café entre outras *commodities*, porém as notícias não são filtradas segundo um contexto específico como impacto em oferta e demanda ou pela perspectiva de IC.. Opfine é gratuito, as demais têm versões gratuitas limitadas.

4 *DESIGN RESEARCH*

4.1 Entendimento do Problema

A investigação teve início pelo conhecimento do problema na primeira etapa do método *Design Science Research*, através de revisão da literatura, apresentada no Capítulo 2, sobre o agronegócio do café, gestão de risco com derivativos em que foi delineado o problema, revisão da literatura sobre os conceitos de Estratégia, Tomada de Decisões, *Business Intelligence* e Inteligência Competitiva, pelo caráter do problema no contexto administrativo, e tecnologias envolvidas em sistemas baseados em mineração textual para Inteligência Competitiva.

Os temas estudados na revisão guiaram o método de observação participante, entrevistas e reuniões com os especialistas do Bureau de Inteligência do Café, a partir das quais foram definidos requisitos de IC para a cafeicultura e quais informações a serem coletadas da *web* para atender estes requisitos, respondendo as questões propostas 1 e 2. Os especialistas participaram também como usuários finais na construção do artefato. Este capítulo apresenta o perfil dos participantes, a descrição dos dados disponíveis, os passos e resultados da primeira etapa realizada em dois ciclos do método *Design Science Research.*

4.1.1 Perfil dos participantes

Três pesquisadores do Centro de Inteligência em Mercados estiveram envolvidos no processo de análise e avaliação de notícias e dois deles também profissionais do CIM, participaram na definição dos requisitos de inteligência e do sistema. A Tabela 4 apresenta a descrição de perfil dos participantes.

Tabela 4 – Perfil dos Participantes

Formação Acadêmica	Cargo no CIM	Qualificação Profissional e Experiência
Engenheiro Agrônomo e Mestre em Agronomia/Fitotecnia (UFLA)	Há sete anos Coordenador de Pesquisas e Serviços em Gestão. Principal projeto: Campo Futuro CNA/UFLA/CIM	Analista de mercado de café desde 2009 coordenador de gerenciamento de custos e comercialização para 12 agronegócios (Cafeicultura e Fruticultura) em 13 estados do Brasil
Doutorando em Administração e Mestre em Administração (UFLA). Graduado como Tecnólogo em Cafeicultura (IF Sul de Minas)	Há 6 anos Coordenador do Bureau de Inteligência Competitiva do Café	Oito anos de experiência em análises relacionadas ao setor cafeeiro, com ênfase em inteligência competitiva e certificação. Autor de artigos científicos e de opinião sobre o setor.
Graduado em Economia (FEA/USP) e Mestre em Modelagem Matemática Financeira (FEA-IME/USP)	Pesquisados no CIM a 4 meses	Economista com 20 anos de experiência em administração de fundos de investimento

A tabela mostra que os participantes têm diferentes formações acadêmicas mas com atuação e experiência no setor cafeeiro e mercado, o que respalda a avaliação das notícias e levantamento de requisitos.

Estes participantes foram selecionados pois são os responsáveis pela análise das notícias para redação do Relatório de Tendências do Café divulgado mensalmente pelo BIC.

4.1.2 Descrição dos dados

As notícias são pesquisadas na *web* pelos especialistas via ferramenta de busca. Aquelas que os especialistas julgam relevantes são armazenadas em um banco de dados que é a base para a produção de relatórios e análise para

Inteligência Competitiva do BIC. Também são destacados e armazenados trechos relevantes das notícias.

Duas entidades do banco de dados são utilizadas no trabalho: notícias e trechos. A tabela notícias contém data de publicação, data de coleta, *link* da notícia, título, texto, fonte e categoria que indica o setor da cadeia produtiva para o qual a notícia é relevante. A tabela tópicos contém trechos relevantes de determinada notícia significativos para o estudo.

São 3.112 notícias em inglês de 06/01/2011 a 29/10/2015 e 10.147 trechos. A seguir, um exemplo de notícia rotulada no banco de dados para indústria e em negrito, trechos da notícia armazenados no banco de dados.

Starbucks on Thursday made its first broad push into India, a move to build on its growth in Asia and open the door to new sources of coffee beans and store growth opportunities. *Starbucks said it signed a nonbinding memorandum of understanding with Tata Coffee, which has supplied premium coffee beans to Starbucks. The companies will collaborate on sourcing and roasting high-quality green arabica coffee beans and explore development of Starbucks retail stores in Tata retail outlets and hotels Tata Coffee, one of India's largest growers and exporters of coffee, is part of India's Tata Group, a huge conglomerate that makes everything from Jaguar cars to Eight O'Clock coffee and Good Eart tea.* ***In addition, Starbucks and Tata said they will consider investments in facilities to export coffee from India to other countries. Tata owns 19 coffee estates in southern India.***

4.1.3 Observação participante e entrevista livre

A pesquisa foi realizada no Bureau onde o pesquisador participou de todos os processos do sistema de coleta e análise de notícias para escrita dos relatórios. A interação com o grupo participante culminava em reuniões periódicas. As entrevistas foram conduzidas nessas reuniões *in loco*, com a

presença de todos os participantes. O processo pode ser dividido em momentos-chave que produziram os resultados desta etapa, descritos a seguir.

4.1.3.1 Definição do problema

Para a definição do problema apresentado no Capítulo 1, em reunião com os especialistas do Bureau, foi considerado seu objetivo de oferecer informações e análises que contribuam para planejamento e tomada de decisões pelos agentes da cadeia agroindustrial do café. Este objetivo é amplo e inclui diferentes agentes (empresas, cooperativas, produtores rurais) com objetivos distintos, mas em comum a exposição à variação de preço do café.

Nesta linha, dois pontos foram apresentados como estratégicos, para as empresas do setor, em nível de estratégia funcional. O primeiro é identificar o momento favorável de realizar *hedge* no mercado futuro de café para gerenciamento de risco e em nível de estratégia de negócios e o segundo é identificar eventos que afetam os fundamentos do setor cafeeiro e competitividade do produto. Este último também estratégico para o país enquanto maior produtor mundial de café Arábica.

Existem, nas notícias coletadas pelo BIC, eventos que impactam diferentes setores da cadeia produtiva do café. Estes eventos conduzem a análise qualitativa dos especialistas para a escrita dos relatórios, entretanto não há conhecimento suficiente sobre a existência de relação entre estes eventos e a variação de preço que permita tomar decisões estratégicas para o gerenciamento de risco. Existem também eventos relacionados à competitividade, mas não estão organizados em um modelo formal que organize as informações presentes nas notícias para gerar Inteligência Competitiva.

Portanto, duas necessidades foram apontadas pelos especialistas: verificar se a ocorrência de eventos tem relação com a variação de preço do café de forma adequada para a decisão de gerenciamento de risco e as ferramentas

para obter Inteligência Competitiva a partir das notícias que apoiem a tomada de decisão na cafeicultura.

Durante a observação participante também foi constatado que o processo de coleta de notícias da *web* é restrito à capacidade de leitura e tempo dos especialistas, o que leva a outra demanda do ponto de vista técnico confirmada pelos participantes: a possibilidade de coleta e classificação automática de notícias.

Esta interação motivou a busca por processos de Inteligência Competitiva e modelos de sistemas baseados em mineração *web* e textual. As etapas do modelo de processo adotado nesta pesquisa (PELLISSIER; NENZHELELE, 2013b), explicado no Capítulo 2, fundamentaram as questões específicas como pauta da próxima reunião.

4.1.3.2 Requisitos de inteligência competitiva

Nesta interação, a primeira questão do processo de IC foi apresentada: **Questão 1 –** Quais os requisitos de inteligência para gerenciamento de risco e competitividade na cafeicultura?

A partir de entrevista livre com cada especialista, foram identificados Tópicos Fundamentais de Inteligência para as necessidades de estratégia, levantadas na interação anterior, que possibilitem a obtenção de inteligência acionável. Antes porém, foi necessário entender e configurar as forças competitivas no ambiente da cafeicultura a ser monitorado sob a perspectiva do país. Foi apresentado e adotado o modelo de Forças Competitivas de Porter adaptado, através do qual, para o contexto do mercado de café e análises do Bureau, foram identificadas três partes no ambiente.

a) Rivais: Países concorrentes produtores de café;

b) Compradores: Países importadores, compradores de café e empresas compradoras de café;
c) Fornecedores: Produtores rurais de café arábica, robusta e diferenciado.

Além dos agentes, foi considerada a influência de oferta e demanda no preço do café, atributo comum a todos os agentes e característica que suscita o conhecimento sobre variação da produção em países concorrentes, variação de consumo em países compradores, ameaças a fornecedores e outros fatores que impactam oferta e demanda.

A partir desta configuração, foram identificados os Tópicos Fundamentais de Inteligência do tipo alerta antecipado (sinais de mercado), apresentados em forma de perguntas e assuntos relacionados:

a) Há tendência de aumento de volatilidade do preço do café no mercado?;
b) Há tendência de aumento de produção em países rivais? Indicadores que podem provocar o aumento de produção: Uso de tecnologias, treinamentos, incentivos governamentais e novas áreas de plantio;
c) Como está o desenvolvimento da indústria em países rivais? Investimento em novas fábricas, produtos como cafés solúveis e instantâneos;
d) Como as empresas (compradores) atuam em países rivais? Suporte à atividade no país, investimento, treinamento, desenvolvimento de técnicas e plantação;
e) Como as empresas (compradores) atuam em países compradores? Expansão, abertura de novas lojas, lançamento de novos produtos, bebidas e equipamentos;

f) Quais as estratégias de países rivais? Investimento em cafés especiais e certificados, investimento em sustentabilidade, selos de qualidade, cafés orgânicos, *gourmets* e resistentes;
g) Incidência de pragas e doenças do café em países rivais;
h) Possíveis rupturas em fornecedores. Quebras de safra por fatores climáticos, pragas e doenças.

Segundo os especialistas do BIC, esses tópicos são relevantes para o estudo de Inteligência Competitiva na Cafeicultura.

4.1.3.3 Coleta de informações

Nesta interação a questão referente à segunda etapa do processo de IC é: **Questão 2 –** Quais informações devem ser coletadas da *web* para atender os requisitos da Questão 1?

Os dados para o Bureau são as notícias publicadas na *web*. A transformação em informação começa pela seleção e classificação manual de notícias relevantes para análise segundo julgamento dos especialistas. Os membros do Bureau utilizam ferramentas de busca para pesquisar diariamente as notícias na *web*, utilizando uma lista de termos definida pelos profissionais. Esta lista inclui nomes de empresas – Starbucks, Nestle, Green Mountain, Lavazza, dentre outras, países, e termos comuns na cafeicultura: *coffee*, *drink*, *coffee shop*, *production*, etc. Durante a pesquisa, à medida que a resposta para a primeira questão se delineava juntamente com a possiblidade de automatizar o processo, surgiu a necessidade de formalizar e ampliar a busca em duas dimensões: ambiente e agentes.

O ambiente é entendido como a cadeia produtiva do café e fatores que a influenciam, e os agentes foram os definidos anteriormente. Portanto para atender os requisitos da questão anterior devem ser coletadas notícias referentes

aos tópicos definidos e sobre os segmentos da cadeia produtiva: de fornecedores de insumos, máquinas e equipamentos a varejo nacional e internacional. Como rivais: África, Ásia, Buon Ma Thuot, China, Colômbia, Etiópia, Guatemala, Honduras, Índia, Indonésia, Quênia, México, Peru, Tanzânia, Uganda e Vietnã.

Como países compradores: Alemanha, Austrália, Canada, China, Coreia do Sul, Emirados Árabes, Espanha, Estados Unidos, Filipinas, Finlândia, França, Inglaterra, Itália, Japão, Portugal, Reino Unido, Rússia, Suécia e Suíça.

Empresas compradoras: Baskin-Robbins, Biggby Coffee, Blenz Coffee, Caffè Nero, Caribou Coffee, Coca-Cola, Coffee Republic, Costa Coffee, Douwe Egberts, Dunkin' Donuts, Folgers, Gloria Jean's Coffees, Keurig Green Mountain, Kraft, Krispy Kreme, Lavazza, Marley, McDonald's, Nestlé, Sara Lee, Starbucks, Subway, Tata, Testarossa Coffe, Tim Hortons, The Coffee Bean, Tully's Coffee, Coffee Club, Juan Valdez, Alterra Coffee Roasters, Eight O'Clock, Massimo Zanetti, Peet's Coffee and Tea.

Como fatores que influenciam a cadeia produtiva: notícias sobre o clima, pragas e doenças.

Após esta etapa, foi construído o primeiro artefato, apresentado no Capítulo 4.

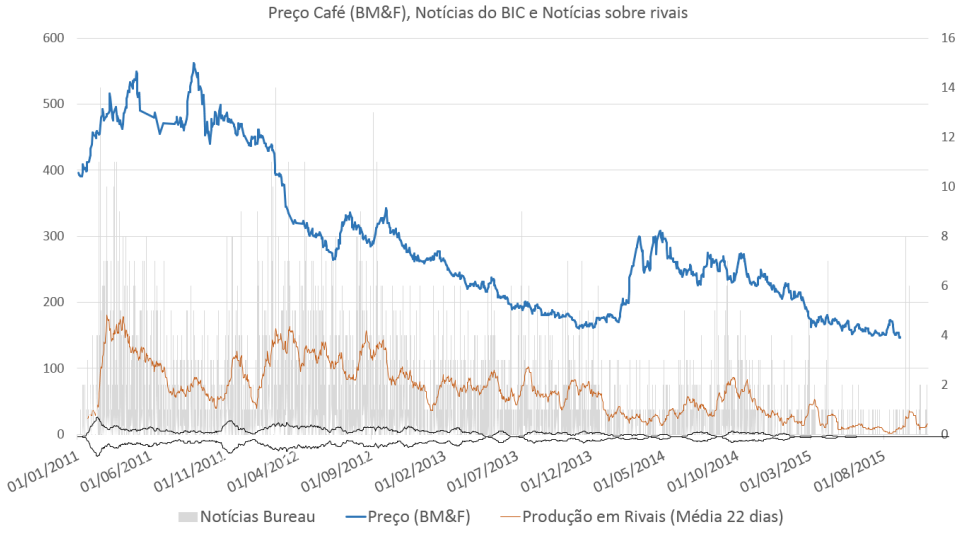
4.1.3.4 Modelos para análise da informação

O primeiro artefato proposto, descrito no Capítulo 4, avançou o conhecimento sobre a coleta e classificação de dados, e impulsionou o retorno à primeira etapa do método, Entendimento do Problema, para uma revisão sobre as ferramentas de IC e resposta a **Questão 4 –** Por quais modelos a informação deve ser interpretada e analisada para IC?

As informações para atender os requisitos de inteligência são eventos (notícias) em ordem cronológica, portanto a ferramenta adequada para interpretação é *Event Timeline Analysis*, especialmente no caso de estudo da

relação com preço para gerenciamento de risco. A Figura 12 mostra o preço do café na BM&F (linha azul), número de notícias coletadas pelos especialistas por dia (barras cinzas), média de notícias nos últimos 22 dias (linha laranja) e a média dos últimos 22 dias de notícias sobre rivais no período (linhas pretas).

Figura 12 - Eventos citando rivais.



É possível observar pontos com ocorrência de notícias mas não sobre rivais. Este tipo de ferramenta possibilita o estudo de eventos específicos, no caso do Bureau, as ocorrências relacionadas aos tópicos de inteligência definidos.

Os resultados alcançados com o primeiro artefato e a revisão das ferramentas para IC indicaram a necessidade de adquirir mais informações específicas para Inteligência Competitiva. Para a análise sobre gerenciamento de risco, é necessário coletar evidências qualitativas que podem impactar a oferta e demanda de café e, por consequência, a variação no preço. A Tabela 5 apresenta títulos de notícias consideradas impactantes para oferta e demanda pelos especialistas.

Tabela 5 - Oferta e Demanda.

Polaridade	Oferta	Demanda
Positivo	*"Colombia coffee production recovers"*	*"Lavazza announces manufacturing facility in India"*
Negativo	*"Drought might harm Uganda's coffee harvest"*	*"Price of coffee may give bigger jolt than caffeine"*

Para interpretações sobre o ambientes e concorrentes, com foco na cafeicultura do país, a análise *SWOT* foi apontada como adequada, portanto, é necessário coletar notícias que indiquem oportunidade, ameaça, força ou fraqueza para o Brasil. A Tabela 6 mostra os títulos de notícias coletadas em março de 2011 pelos especialistas como exemplo de *SWOT* referentes à produção.

Tabela 6 – SWOT.

Strengths	***Weaknesses***
"Coffee Farmers May Get Help to Store Beans"	*"As weather shifts, coffee farmers struggle to protect crops"*
Opportunities	***Threats***
"Colombia rains risk new jolt for coffee prices"	*"India coffee exports surge to 3 lakh tons"*
"Drought might harm Uganda's coffee harvest"	*"Tanzania coffee production to reduce by at least one-quarter"*
"Indonesia supply woes to add froth to coffee market"	*"Vietnam starts futures trading at coffee exchange"*

Por exemplo, o trecho de notícia extraída da *web*: "Baixa produção de café arábica na Colômbia pode favorecer exportação brasileira" é uma evidência qualitativa que representa uma oportunidade para o Brasil e ao mesmo tempo um evento que contribui para a diminuição da oferta. Está relacionada com o tópico de inteligência 1 – Tendências para aumento de produção em países rivais, segmento de produção primária da cadeia produtiva sobre o país rival Colômbia.

4.1.4 Resultados

As propostas resultantes desta etapa foram os requisitos de inteligência para gerenciamento de risco e competitividade na cafeicultura, informações a serem coletadas da *web* e um modelo para analisá-las sob a perspectiva de IC. A partir destas propostas, a próxima etapa da *Design Science Research* é a Sugestão, em que foram testados dois artefatos construídos nesta pesquisa a partir das questões no processo de IC como tentativa de solução ao problema.

4.2 Sugestão

Para projetar e desenvolver um artefato os dados precisam de organização formal em um modelo que capture as características da cafeicultura para gerar informação útil para tomada de decisão. Para os analistas do Bureau, as notícias extraídas da *web* estão relacionadas a diferentes dimensões do mercado com fatos relevantes que impactam a variação de oferta e demanda de café no mundo.

Entretanto, a relevância é específica para determinado setor, por exemplo, um fato sobre incentivo fiscal para plantio de café em determinado país é considerado relevante para a produção e é analisado pelo especialista nesta dimensão. Ao passo que notícias sobre determinada rede de cafeterias é analisada em um contexto diferente. Desta forma, a cadeia produtiva do café é usada como referência na coleta de notícias pelos especialistas que definiram categorias para organizar a recuperação e classificação de informações extraídas da *web*, ao passo que através dela é possível delimitar dimensões para análise setorial. As categorias são apresentadas na Tabela 7.

Tabela 7 - Categorias da Cadeia Produtiva do Café.

Categoria	Setores	Segmentos
Indústria	Fornecedores de Insumos, Máquinas e Equipamentos	Indústria de Máquinas e Implementos, Produtores de Mudas, Indústria de Defensivos e Fertilizantes
Produção	Produção Primária	Produtores de Café Robusta, Produtores de Café Arábica, Produtores de Café Diferenciado
	Primeiro Processamento	Cooperativas, Maquinistas
	Segundo Processamento	Empresas de Solúvel Nacionais, Empresas Torrefadoras Nacionais, Cooperativas
Bebidas	Vendedores Nacionais	Exportadores, Cooperativas e Central de Cooperativas
	Compradores Internacionais	Empresas de Solúvel (Internacional), Indústria de Soft-drinks, Empacotadores de produtos de solúvel, Empresas de Torrefação (Internacional)
Cafeterias	Varejo Nacional e Internacional	Vending Machines, Mercado Institucional, Lojas de Café Pequeno. Varejo, Supermercados, Bares e Restaurantes

Nesta linha, foram definidas categorias de fatos que podem indicar impacto na oferta e demanda e/ou indicar oportunidades, ameaças, forças e fraquezas. Descritos na Tabela 8 em categorias.

Tabela 8 - Categoria de Fatos. (Continua)

Categoria	Descrição	Exemplo
Produção e Exportação	Informação sobre aumento ou diminuição de produção e exportação.	*"Colombia coffee production grows 22%, exports by 23%"*
		"Nestle Nespresso to boost coffee production in Africa"
Eventos naturais	Incidência de doenças e pragas, variações climáticas que favorecem ou comprometem a atividade.	*"Significant amounts of rain interrupted the harvesting process last month as the wet weather obstructed harvesting and drying of the beans."*
		"The unfavorable weather conditions (heavy rains) in Africa damaged the coffee crop."

Tabela 8 - Categoria de Fatos. (Conclusão)

Categoria	Descrição	Exemplo
Políticas	Incentivos governamentais, ações do governo.	*"Uganda to get governmental help for coffee growers"*
Indicação Geográfica	Investimentos na produção de cafés de qualidade e certificados.	*"Sara Lee announces five-year sustainable coffee plan"* *"McCafes in Brazil Serving 100% Rainforest Alliance Certified Coffee"*
Expansão de Indústria	Ações que indicam expansão e abertura de novos mercados. Abertura de novas fábricas. Parcerias entre o setor privado e países produtores.	*"Nestlé expands its largest soluble coffee factory in Europe"* *"Tipton Mills launches first ever probiotic*
Expansão de Cafeterias	Ações que indicam expansão e abertura de novos mercados. Abertura de novas lojas. Parcerias entre o setor privado e países produtores.	*"Marley Coffee Expands Distribution in Chile"* *"Starbucks Will Open 1550 New Stores In 2014"*
Consumo	Indicam variações no consumo	*"o consumo por grãos gourmet e especiais mais que duplicou, comparando-se 2008 com o ano passado, foi de 1,2%, para 2,7%."*
Sustentabilidade	Eventos, ações, iniciativas voltadas para sustentabilidade da Cafeicultura	*"The Coffee Climate Care Project adresses climate change in the Vietnamese coffee"* *"SAP and the Colombian Coffee Growers Federation bolster sustainable coffee farmi"*
Empresas em Produtores	Investimento feito por empresas em países produtores	*"Nestle supports coffee farmers in India as demand for Nescafé grows"* *"Coffee Giants target Chinas Yunnan plantations"*

Tabela 8 - Categoria de Fatos. (Conclusão)

Categoria	Descrição	Exemplo
Pesquisa	Pesquisas para avançar a Cafeicultura	*"Panamá: Promecafé lança proposta para melhorar genética do café"*
		"Cafeicultores da BA terão estação experimental de café conilon"
Café Especial	Investimento, produção e exportação de cafés especiais	*"Brazilians, Colombians acquire taste for goumet coffee"*
		"The Colombian Coffee Growers Federation exports record number of specialty coffee"
Especulação	Fatores que podem provocar volatilidade no preço por movimentos especulativos	*"National coffee production rise 9%"*
		"Robusta coffee going the arábica way, prices fall below cost of production on go"

Estes modelos de categorias permitem a classificação para a interpretação e análise com as ferramentas para IC propostas na etapa anterior. Assim em resposta à **Questão 3** – Como as informações devem ser organizadas? (3. Triagem, captura e armazenamento de informações), as notícias têm os atributos descritos na Tabela 9.

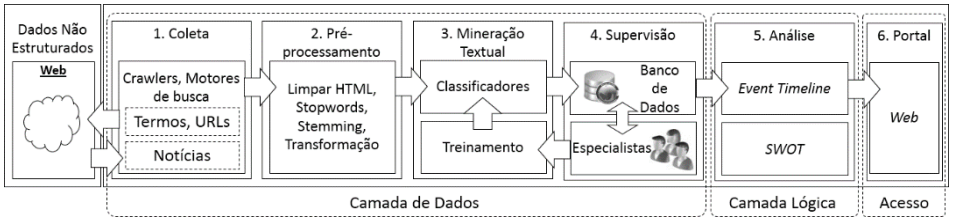
Tabela 9 - Atributos das notícias.

Atributo	Descrição
Data de Inserção	Data em que a notícia foi inserida no banco de dados
URL	Endereço eletrônico da notícia
Data de Publicação	Data em que a notícia foi publicada na *web*
Palavra Chave	Termo pesquisado que retornou a notícia (caso use motores de busca)
Fonte	Agência que publicou a notícia
Título	Título da notícia
Texto	Texto da notícia, a notícia propriamente dita
Categoria da Cadeia Produtiva	Nível da cadeia produtiva ao qual a notícia está relacionada (TABELA 7)
Categoria do Evento	Fato relevante ao qual a notícia se refere (TABELA 8)
SWOT	Se a notícia representa força, fraqueza, oportunidade ou ameaça
Impacto de Curto Prazo	Onde a notícia impacta: oferta ou demanda
Polaridade e Intensidade de CP	Quanto impacta e se é positivo, negativo ou neutro no curto prazo
Impacto de Longo Prazo	Onde a notícia impacta: oferta ou demanda
Polaridade e Intensidade de LP	Quanto impacta e se é positivo, negativo ou neutro no longo prazo

Os atributos não sublinhados são coletados da *web* e extraídos da notícia. Os atributos sublinhados são classificados automaticamente pelo sistema baseado em Mineração Textual por classificação supervisionada.

O projeto para construção do artefato seguiu processos de mineração de dados com as etapas até análise, ilustrados na Figura 13, conforme estudo dos modelos tradicionais para extração de conhecimento (MORAES; VALIATI; GAVIAO NETO, 2013; BRAMER, 2013), *Business Intelligence* (BAARS; KEMPER, 2008; BOSE, 2009; WU, 2010) e Sistemas baseados em Mineração Textual para Inteligência Competitiva (DAI; KAKKONEN; SUTINEN, 2010, 2011a, 2013).

Figura 13 - Arquitetura do Sistema.



A Figura 13 mostra o sistema dividido em três camadas principais, camada de dados, onde acontecem as tarefas de coleta, pré-processamento, classificação e organização do dado para a análise; a camada lógica com ferramentas para análise sobre IC e a camada de acesso, por um portal *web*. As camadas possuem módulos responsáveis pelas tarefas, descritas a seguir.

4.2.1 Coleta

Os dados são coletados da *web* por duas vias: rastreadores (*web crawlers*) e consulta com motores de busca. Os rastreadores recebem como entrada uma lista de endereços da *web* (*URL – Uniform Resource Locator*), visitam cada endereço extraindo o texto da página e identificando novos *links* para visitar recursivamente de acordo com regras pré-definidas.

A lista de entrada (semente), para o sistema, deve conter *sites* de agências especializadas em notícias sobre o café para evitar a inserção de notícias irrelevantes como ruído no sistema, comprometendo a análise. Os endereços definidos como sementes para sistema são listados e atualizados pelos especialistas.

A vantagem do uso de rastreadores é a independência de tecnologia proprietária para busca e a flexibilidade para navegação na *web* sem restrições de requisições ou limite de acesso. Entretanto, o desafio passa a ser determinar a relevância do conteúdo visitado.

O uso de motores de busca tem como entrada um termo para pesquisa. Procura o termo em documentos da *web* indexados por critérios de relevância próprios, e retornam os *links* com a URL destes documentos. Existem ferramentas que permitem selecionar a busca na web pelo termo especificamente em notícias, com resultado ordenado por relevância, por data ou em um período específico.

A vantagem é ter uma ferramenta de busca amplamente utilizada e aprimorada especificamente para este fim por empresas do mercado. Porém é necessário observar as restrições de acesso e as limitações de requisições e uso. É importante também verificar a ocorrência de *links* indisponíveis e *sites* que exigem autenticação para leitura da notícia.

Na abordagem com motores de busca, para aprimorar a pesquisa por notícias na *web* com resultados mais abrangentes e relevantes, foi definido um dicionário de termos comuns ao contexto estudado. Os termos foram definidos juntamente com os especialistas e pelo estudo da base de dados existente.

A definição com especialistas considerou os requisitos de inteligência, cadeia produtiva, forças competitivas e fatores que impactam oferta e demanda. O dicionário tem as palavras selecionadas conforme Tabela 10.

Tabela 10 – Descrição do Conjunto de Palavras para Pesquisa na Web.

Descrição do Conjunto de Palavras	Tópicos de Inteligência Relacionados
Empresas: com o nome dos principais agentes do mercado, produtores, consumidores e torrefadores. Exemplo: Nestle, Lavazza, Starbucks, entre outros.	3 e 4
Países: nomes dos países produtores e compradores de café.	1, 2, 3, 4, 5 e 6
Clima: termos como *rain*, *storm*, *dry*, *drought* e situações que impactam a safra.	7 e 8
Doenças e Pragas: nomes de doenças e pragas que afetam o cultivo de café.	6 e 7
Qualidade: termos relacionado a certificações e selos de qualidade.	2 e 5
Relação empresas países: palavras que indiquem ações e relação de empresas em países rivais e compradores – *support, training, investment, open, expand, sales, coffee machines.*	3 e 4
Produção: palavras relacionadas a produção de café em países rivais – *raise, increase, new areas, government, production*.	1, 6, 7 e 8
Indústria: palavras relacionadas a indústria – *instant, soluble, new factory*.	9

A Tabela 10 apresenta a descrição do conjunto de palavras para consulta e uma coluna com os tópicos de inteligência apresentados na Seção 4.3.2 aos quais os termos estão relacionados.

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 e relevância da própria notícia para o contexto do sistema – cafeicultura, problemas abordados nos módulos seguintes.

4.2.2 Pré-processamento

Pré-processamento é o processo de preparação do texto que promove o desempenho da classificação textual (HADDI; LIU; SHI, 2013). O módulo de pré-processamento tem dois processos: limpeza e transformação. A limpeza tem como objetivo preparar o conteúdo apenas com o texto essencial da notícia, assim inclui tarefas em nível sintático: retirada de *tags HTML*, *scripts* e outros

ruídos presentes em texto extraído da *web* irrelevantes para o contexto. Outra tarefa aplicada é a remoção de *stopwords* – palavras que não são significativas para representar uma categoria e aparecem com frequência como artigos, pronomes, preposições e advérbios.

Para reduzir a variação dos termos, por exemplo: *producer*, *producers* e *produce*, *producing*, para uma única representação, aplica-se o procedimento denominado *stemming* (WEISS et al., 2010).

Na transformação, o conteúdo é convertido em uma representação matemática adequada para a entrada de um classificador que define uma categoria para a notícia pela comparação estatística do texto pré-processado com uma base previamente definida – base de treinamento do classificador. Procedimento executado no módulo seguinte, descrito na próxima seção.

4.2.3 Mineração textual: treinamento e classificação

A abordagem deste módulo é a classificação supervisionada por aprendizado de máquina. Para isso, no módulo de Supervisão, os especialistas classificam manualmente as notícias que selecionam da *web*, criando assim um conjunto de notícias rotuladas usado como base para treinamento de algoritmos classificadores de notícias coletadas da *web* pelo sistema.

Um documento é usualmente representado como "*bag of words*" (SEBASTIANI, 2002) e transformado em uma representação numérica adequada para entrada de cada classificador. Grande volume de dados apresentam ruídos, como palavras irrelevantes, que comprometem o desempenho dos classificadores, portanto, exigem o processo de *feature selection* para reduzir a dimensão dos dados. O objetivo é remover feições (termos, palavras ou frases) irrelevantes e considerar um conjunto mínimo capaz de atingir uma classificação comparável a que usa todo o conjunto.

O treinamento, portanto, consiste em gerar um modelo de representação numérica do texto da base de notícias classificadas pelos especialistas. A representação utilizada é *TF-IDF* (*Term Frequency – Inverse Document Frequency*) que considera a frequência de cada termo no texto de uma notícia e sua relevância para todo o conjunto de notícias. Conforme Manning, Raghavan e Schützel (2008):

$$TF - IDF_{t,d} = TF_{t,d} \times IDF_t, \qquad (2)$$

Em que:

$IDF_t = \log\frac{N}{DF_t}$;

$TF_{t,d}$ = número de ocorrências do termo t no documento d;

N = número de documentos no conjunto;

DF_t = número de documentos no conjunto que contém o termo t.

As representações numéricas de cada texto são associadas as suas respectivas categorias armazenadas no banco de dados pela classificação dos especialistas segundo os atributos pré-definidos. A partir deste modelo, as notícias extraídas da *web* poderão ser classificadas no módulo de Classificação.

Nesta etapa, um método de classificação determina a classe de uma instância não classificada com base no modelo criado na fase de treinamento.

Um classificador é uma função f que mapeia vetores de características $x \in X$ em classes rotuladas $\in \{1, ..., C\}$, onde X é o espaço de características. Os rótulos são desordenados (categorias) e mutuamente exclusivos. A meta é aprender f de um conjunto rotulado de treinamento com N pares de entrada – saída (x_n, y_n), n = 1 : N. Este é considerado aprendizado supervisionado.

Existem vários métodos para esta tarefa, dentre eles: Naïve Bayes (JOHN; LANGLEY, 1995), Naïve Bayes Multinomial (MCCALLUM; NIGAM,

1998), Complement Naïve Bayes (RENNIE et al., 2003), Discriminative Multinomial Naïve Bayes (SU et al., 2008), J48 (QUINLAN, 1993), Random Tree e Support Vector Machines (PLATT, 1998).

O resultado deste módulo são notícias classificadas segundo critérios pré-definidos pelos especialistas para a análise.

4.2.4 Supervisão

O objetivo do módulo de Supervisão é manter a integridade e atualização da base de treinamento. Para isso, permite que os especialistas insiram novas notícias e avaliem a coleta e classificação automática realizada pelo sistema através de uma interface com o banco de dados.

Antes de iniciar o primeiro módulo, de coleta, é necessário uma base de dados rotulada pelos especialistas. A base de dados rotulada de acordo com a cadeia produtiva estava em construção desde o começo do projeto, portanto foi necessário revisar as categorias e algumas notícias com categorias não pertinentes ao estudo.

Após a conclusão do primeiro ciclo do método, verificada a necessidade de informação para Inteligência Competitiva, proposto acrescentar ao módulo de supervisão uma interface para classificação manual das notícias quanto à oferta, demanda e SWOT.

4.2.5 Análise

Este módulo atua na informação organizada para IC – notícias coletadas e classificadas do banco de dados pelos módulos anteriores. A partir delas é possível gerar relações e visualizações que promovam a análise.

Pela revisão da literatura e conversa com os especialistas, foram definidas funcionalidades do sistema para o módulo de análise que atendam aos requisitos de IC pré-definidos. Para cada funcionalidade, um conjunto de

notícias deve ser recuperado e organizado para permitir inferências sobre os requisitos de IC.

A Tabela 11 apresenta as funcionalidades do sistema e notícias de entrada de acordo com a classificação para Dimensão da Cadeia, Fato Relevante, Oferta e Demanda (O/D) e SWOT.

Tabela 11 - Funcionalidades do Sistema.

Forças	**Funcionalidades**	**Dimensão da Cadeia**	**Fato relevante**	**O/D**	**SWOT**
Rivais	Identificar aumento de produção	Produção	Produção e Exportação Políticas Eventos Naturais Pesquisa Café Especial	O+	T
	Detectar desenvolvimento da indústria	Indústria	Expansão da Industria	D+	T
	Identificar atuação das empresas	Cafeterias	Expansão de Cafeterias Consumo	D+	T
	Monitorar estratégias	Indústria Produção Bebidas	Sustentabilidade Indicação geográfica Café Especial Pesquisa	OD	T
Compradores	Identificar a incidência de pragas e doenças	Produção	Eventos Naturais	O-	O
	Identificar atuação em países compradores	Indústria Cafeterias	Empresas em Compradores Café Especial Indicação Geográfica Expansão de Cafeterias	-	O
Fornecedores	Identificar rupturas em fornecedores	Produção	Eventos Naturais	O-	S T
Ambiente	Detectar eminencia de volatilidade no preço	Todas	Todos	O/D +/-	T

Legenda: O/D – Oferta/Demanda, O+ – Oferta Positiva, O- – Oferta Negativa, D+ – Demanda Positiva, D- – Demanda Negativa, OD – Oferta e Demanda Positivas ou Negativas.

Por exemplo, para a funcionalidade: Aumento de Produção em Rivais, é possível recuperar para análise, as notícias da produção, que seja fato relevante

relacionado à Produção e Exportação, Política, Eventos Naturais, Pesquisa e Café Especial, que tenham impacto positivo para oferta, que representem uma ameaça e que possuam nome de algum rival no título ou texto, ou um rival específico, ou outra regra de ocorrência de termo definida pelo analista.

Este conjunto de notícias tem as características que podem indicar o aumento de produção pelo uso de tecnologias, treinamentos, incentivos governamentais, novas áreas de plantio, dentre outros, de acordo com a base de treinamento criada pelos especialistas.

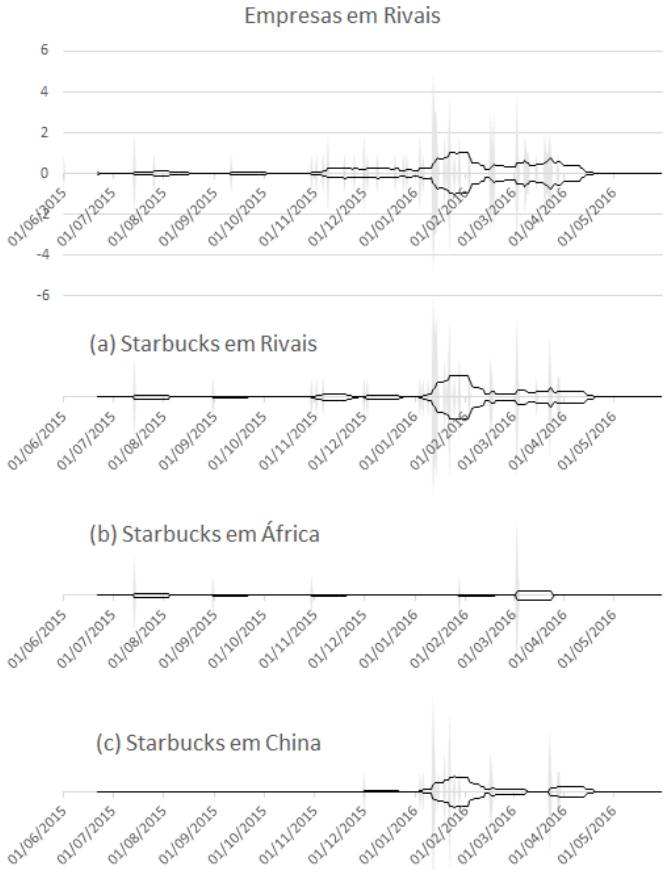
Outro exemplo, para a funcionalidade: Identificar a Atuação das Empresas em Rivais, são selecionadas notícias da classe cafeterias, que sejam fatos para Expansão de Cafeterias e Consumo, que contribuam positivamente para a demanda, representam ameaça e tem como ocorrência o nome de uma das empresas ou alguma delas e um país específico ou algum país, dependendo da análise desejada.

Para completar a funcionalidade e como resposta a **Questão 5 –** Como a inteligência deve ser apresentada para apoiar a tomada de decisões (5. Difusão de Inteligência), revisando a literatura e em reunião com especialistas, foram definidas duas formas de visualizar a informação para análise: Ocorrência de eventos no tempo (*Event Timeline*) e matriz *SWOT*.

4.2.5.1 Eventos

A Figura 14 apresenta o resultado da funcionalidade Identificar a Atuação de Empresas em Rivais, com foco na empresa Starbucks.

Figura 14 - Empresas em Rivais.

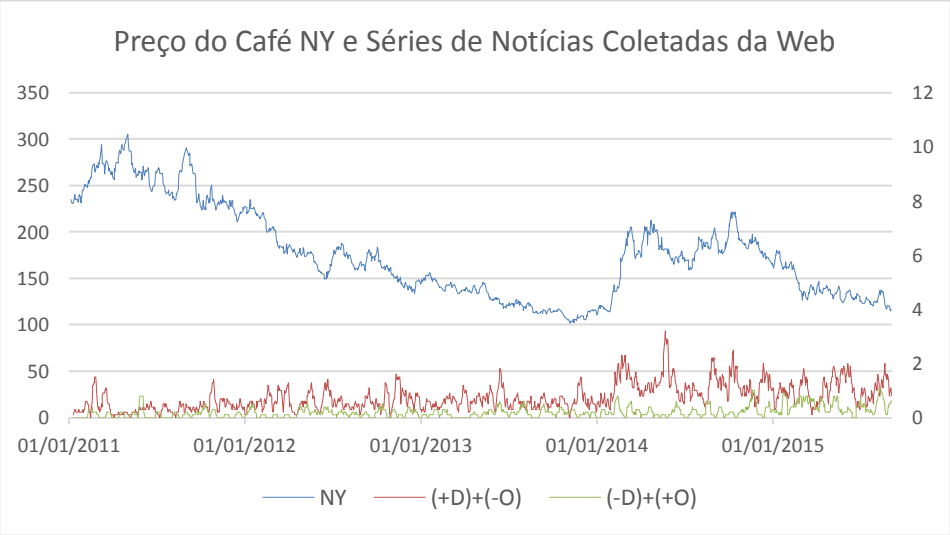


É possível observar na Figura 14 um aumento de notícias pela funcionalidade: Identificar Atuação de Empresas em Rivais, em janeiro e fevereiro de 2016. A partir desta observação, o especialista pode direcionar a análise para este requisito. Ou observar uma empresa específica nos rivais, Figura 14 (a), ou em um país específico, Figura 14 (b) e (c).

O requisito volatilidade do preço do café no mercado exige uma visão dos fatores que impactam oferta e demanda em relação ao preço. Neste sentido, a proposta é um gráfico que inclua as variáveis ocorrência de notícias que

impactam positivamente a oferta e demanda (contribuem para aumento) e negativamente (contribuem para diminuição) em paralelo ao preço, Figura 15.

Figura 15 – Séries de Notícias Coletadas da Web e Preço.



A Figura 15 apresenta o preço do café na bolsa de Nova York (NY), média de 10 dias de ocorrências de notícias coletadas da web que impactam positivamente a demanda mais negativamente oferta: (+D)+(-O), e negativamente a demanda mais positivamente a oferta: (-D)+(+O).

4.2.5.2 Swot

Para a análise, é importante reconhecer fatores internos e externos que auxiliam as decisões estratégicas na cafeicultura em geral. Com o foco no país, a análise considera notícias que apontam fatores internos que representam fraqueza ou força, e fatores externos que representam ameaças ou oportunidades.

Neste sentido, foram propostas duas formas de visualização das notícias que facilitam a construção de uma matriz SWOT e a análise de oportunidades e ameaças em paralelo às funcionalidades anteriores: Quadro com as notícias que

indicam fraqueza, força, oportunidades e ameaças em cada setor analisado e um gráfico com a soma das ocorrências dessas notícias em período selecionado pelo especialista para análise.

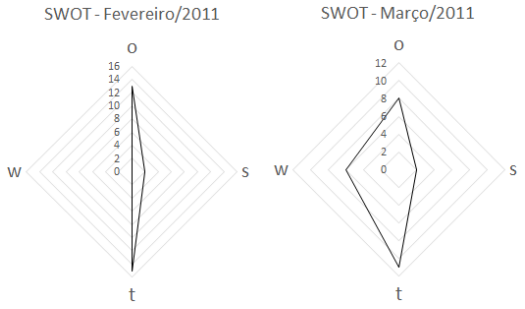
A Tabela 12 apresenta notícias publicadas em março de 2011, classificadas como Produção e organizadas de acordo com a classificação para *SWOT*.

Tabela 12 - Notícias Organizadas por SWOT.

Oportunidades	**Forças**
"Indonesia supply woes to add froth to coffee market"	*"Why the rising cost of coffee isn't a bad thing"*
"Coffee exports may fall in Apr-June"	*"Coffee Farmers May Get Help to Store Beans"*
"Drought might harm Uganda's coffee harvest"	
"Coffee Market in India 2011"	
"Blaming Climate Change Won't Help Costa Rica Coffee"	
"Tanzania coffee production to reduce by at least one-quarter"	
"New coffee export rules may slow exports"	
"Colombia rains risk new jolt for coffee prices"	
Ameaças	**Fraquezas**
"Zambia plans to revive its largest coffee producer"	*"Rains lift world coffee prices"*
"Coffee exports up by 47% in January-February"	*"As weather shifts, coffee farmers struggle to protect crops"*
"The need of friendly policies for Uganda coffee farmers"	*"Brazil's Coffee Output May Fall on Weather, Fungus"*
"Colombia and Venezuela seek binational plan for coffee producers"	*"Coffee output could rise faster than some believe"*
"Indonesia Coffee Sales Accelerate on London Prices"	*"Price of coffee may give bigger jolt than caffeine"*
"Vietnam starts futures trading at coffee exchange"	*"For coffee lovers, a serious setback"*
"Vietnam coffee needs stronger link to farmers"	
"Number of Tanzanians drinking coffee is very small"	
"Coffee Exporters Unveil Strategy to Increase Internal Market Share"	
"India coffee exports surge to 3 lakh tons"	
"Tanzania Coffee board assures growers of higher prices"	

A Figura 16 apresenta a distribuição das ocorrências de notícias separadas de acordo com os fatores SWOT em gráfico radial, dos meses de fevereiro e março de 2011.

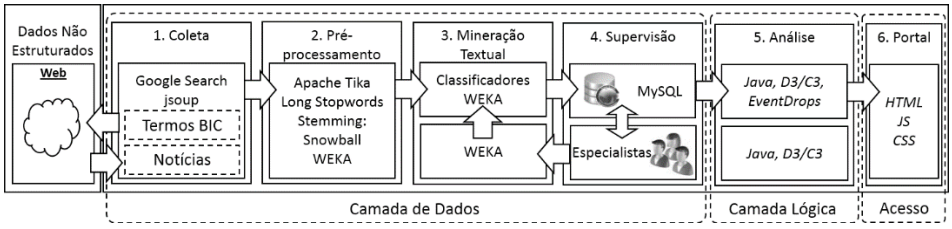
Figura 16 – Gráfico Radial de Notícias Classificadas para SWOT.



4.3 Desenvolvimento

O sistema foi desenvolvido como uma aplicação Java. O processo de mineração de dados foi realizado com a utilização da interface de programação da ferramenta WEKA (HOLMES; DONKIN; WITTEN, 1994) que contém algoritmos para pré-processamento, transformação e classificação de dados. O desenvolvimento seguem as etapas do modelo apresentado na seção 4. A Figura 17 mostra o sistema com as tecnologias em cada módulo.

Figura 17 - Tecnologias em cada módulo do sistema.



4.3.1 Módulo de coleta

A construção do módulo de coleta adotou motor de busca com pesquisa em períodos específicos.

O sistema monta os termos para pesquisa concatenando a palavra *coffee* com as palavras do dicionário definido pelos especialistas. Este procedimento garante a busca da palavra no contexto do café. Por exemplo, uma das palavras

do dicionário é *production*, consultá-la apenas retorna notícias que têm a palavra em diferentes contextos, enquanto a consulta por *coffee+production* recupera notícias que têm a produção no contexto do café. A Tabela 13, mostra os cinco primeiros de 10 resultados de pesquisas com os termos citados.

Tabela 13 - Resultados de Pesquisas.

Resultados de pesquisa por *production*	**Resultados de pesquisa por coffee+production**
"Staggering shifts to avoid heat affecting coal production"	*"Nestlé invests $84m in South African coffee production plant"*
"Toyota resumes production in Japan post quakes"	*"illycaffè CEO Sees Need for More Coffee Production"*
"Namdeb production down 4% to 400 000"	*"Vietnam - Daklak 2016/17 coffee production seen falling to…"*
"Steven Seagal Plans Production *Venture in Thailand"*	*" Future demand and climate change could make coffee a dri…"*
"TS Solartech to expand production capacity in 2H16"	*"Costa Rica seeks to make coffee production carbon-free"*

Para extrair as informações sobre as notícias resultantes da pesquisa, foi utilizada a biblioteca Java denominada *jsoup* - https://jsoup.org/ devidamente configurada para buscar os elementos textuais.

Para recuperar o texto de uma notícia o sistema visita o *link* retornado pelo motor de busca, coleta todo código HTML da página e entrega para o Módulo de Pré-processamento.

A Tabela 14 mostra uma notícia coletada em pesquisa por *coffee+production*.

Tabela 14 - Notícia Coletada.

Atributo	Valor
Título	Coffee production is expected to go up by 9% in 2013-14
URL	http://articles.economictimes.indiatimes.com/2013-07-23/news/40749565_1_coffee-board-coffee-production-crop-production
Fonte	Economic Times
Data de Publicação	23/07/2013
Descrição	NEW DELHI: Coffee production in the country is expected to increase by 9.05 per cent to 3.47 lakh tonnes in the year 2013-14, on account of normal blossom ...
Trecho de código HTML com o texto relevante	[…]<script src="/modal/js/lazyload-min.js" type="text/javascript"></script> NEW DELHI: Coffee production in the country is expected to increase by 9.05 per cent to 3.47 lakh tonnes in the year 2013-14, on account of normal blossom showers in all major coffee growing areas, according to government data. […]

O texto da notícia está envolvido em marcações para apresentação – código *HTML* – e códigos *Javascript* que não interessam para a análise e prejudicam o processo de classificação. Estes ruídos foram tratados no módulo de pré-processamento.

4.3.2 Módulo de pré-processamento

Na fase de pré-processamento, para limpeza do texto foi adotada a biblioteca de classes em Java, Apache Tika (MATTMANN; ZITTING, 2011), que contém funções específicas para este tipo de tratamento de texto. Este procedimento é necessário antes do treinamento, pois existem no banco de dados notícias copiadas manualmente da *web* pelos especialistas que possuem *tags* HTML e outros elementos irrelevantes que se não forem eliminados, entram no treinamento provocando ruídos que comprometem a classificação.

A lista de *stopwords*, considerada está disponível em http://www.ranks.nl/stopwords. O algoritmo para a tarefa de *stemming* utilizado é o *Snowball* (PORTER, 2001), popularmente utilizado para a língua inglesa.

A Tabela 15 mostra o trecho de código capturado pelo sistema após as tarefas de limpeza, eliminação de palavras (*stopwords*) e *stemming*.

Tabela 15 - Notícia Após Pré-processamento.

Atributo	**Valor**
Título	Coffee production is expected to go up by 9% in 2013-14
URL	http://articles.economictimes.indiatimes.com/2013-07-23/news/40749565_1_coffee-board-coffee-production-crop-production
Fonte	Economic Times
Data de Publicação	23/07/2013
Descrição	NEW DELHI: Coffee production in the country is expected to increase by 9.05 per cent to 3.47 lakh tonnes in the year 2013-14, on account of normal blossom ...
Trecho de código HTML com o texto relevante	[…]<script src="/modal/js/lazyload-min.js" type="text/javascript"></script> NEW DELHI: Coffee production in the country is expected to increase by 9.05 per cent to 3.47 lakh tonnes in the year 2013-14, on account of normal blossom showers in all major coffee growing areas, according to government data. […]
Trecho pré-processado	coffee produc country expect increase lakh tonnes year account normal blossom show all major coffee grow area accord government data

No processo de transformação da notícia em um vetor de palavras, o filtro adequado na ferramenta WEKA é o *StringToWordVector*. Sua função foi converter texto em um conjunto de atributos representando a ocorrência de cada palavra no texto. Este conjunto foi armazenado em um vetor que será 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 Manning, Raghavan e Schütze (2008).

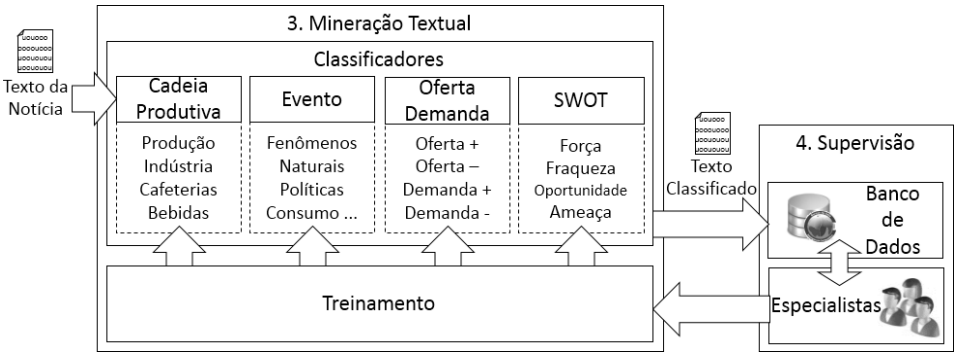
O módulo de pré-processamento também foi acionado no módulo de Mineração Textual para preparar as notícias do banco de dados para treinamento, criando um modelo numérico com o qual as novas notícias foram comparadas para classificação.

4.3.3 Módulo de mineração textual

Seguindo a abordagem de classificação supervisionada, este módulo permitiu a seleção de um conjunto de notícias para treinamento de classificadores.

No caso da classificação textual do sistema proposto, os vetores foram gerados a partir das notícias na fase de transformação (pré-processamento) e as categorias definidas anteriormente. A classificação foi realizada por quatro classificadores diferentes, conforme Figura 18, para identificar os atributos: Categoria da Cadeia Produtiva (TABELA 7), Categoria do Evento (TABELA 8), Impacto, polaridade e intensidade de curto e longo prazo (TABELA 5) e SWOT (TABELA 6).

Figura 18 – Classificadores.



Os métodos mais comuns utilizados para classificação textual são do tipo *Naïve Bayes*, Árvore de Decisão e *Support Vector Machines*, entretanto para fins de descrição do artefato proposto, foi tomado como exemplo o método *Naïve Bayes*.

O método de aprendizado probabilístico da classe *Naïve Bayes* assume que os termos ocorrem independentemente. Dado uma coleção de *N* documentos $\{d_j\}_{j=1}^{N}$ em que cada documento é representado como uma sequência de *T* termos $d_j = \{t_1, t_2, ..., t_T\}$ a probabilidade de um documento d_j ocorrer na classe c_k é dado por:

$$P(c_k|d_j) = P(c_k)\prod_{i=1}^{T} P(t_i|c_k) \qquad (3)$$

Em que $P(t_i|c_k)$ é a probabilidade condicional do termo t_i ocorrer em um documento da classe c_k e $P(c_k)$ é a probabilidade de um documento ocorrer na classe c_k. $P(t_i|c_k)$ e $P(c_k)$ são estimados da base de treinamento. Mais detalhes são encontrados nos trabalhos de Jordan (2002) e Su et al. (2008).

Cada texto de notícia é um documento que terá valores de atributos definidos de acordo com as categorias de cada classificador específico. Ao final a notícia completa é armazenada no banco de dados do BIC. A Tabela 16 apresenta a notícia coletada inserida no banco de dados após a classificação.

Tabela 16 - Notícia Após Classificação.

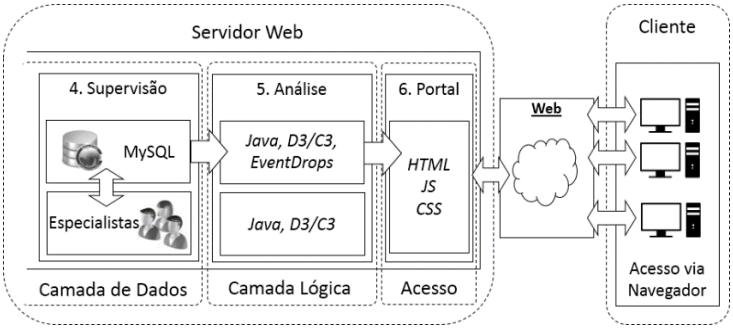
Atributo	Descrição
Data de Inserção	25/04/2016
URL	http://articles.economictimes.indiatimes.com/2013-07-23/news/40749565_1_coffee-board-coffee-production-crop-production
Data de Publicação	23/07/2013
Palavra Chave	coffee+production
Fonte	Economic Times
Título	Coffee production is expected to go up by 9% in 2013-14
Texto	[…] in the country is expected to increase by 9.05 per cent to 3.47 lakh tonnes in the year 2013-14, on account of normal blossom showers in all major coffee growing areas, according to government data. […]
Categoria da Cadeia Produtiva	Produção
Categoria do Evento	Produção e Exportação
SWOT	Ameaça
Impacto de Curto Prazo	Oferta
Polaridade e Intensidade de CP	+1
Impacto de Longo Prazo	Oferta
Polaridade e Intensidade de LP	0

Os atributos Data de Inserção, URL, Data de Publicação, Fonte, Título e Texto foram extraídos da notícia nos módulos de coleta e pré-processamento, os demais atributos tiveram valores atribuídos pelos algoritmos classificadores.

4.3.4 Módulos de supervisão, análise e portal de acesso

Os módulos de Supervisão, Análise e o Portal de Acesso foram desenvolvidos como uma aplicação *web*, conforme Figura 19.

Figura 19 - Arquitetura do Sistema *Web*.

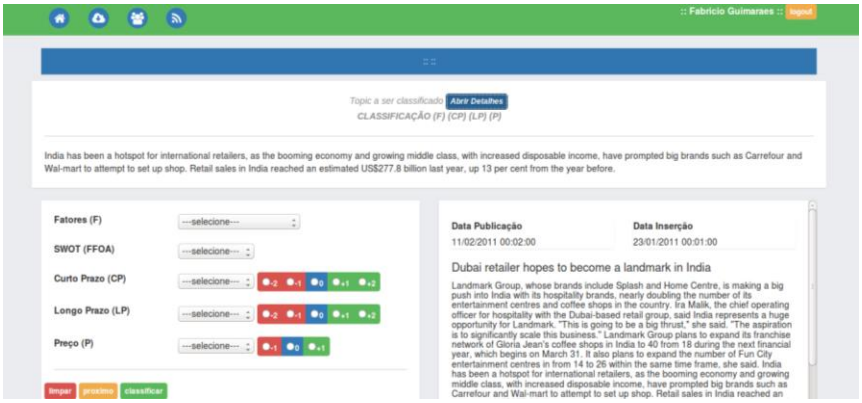


A Figura 19 ilustra uma arquitetura para aplicação *web* com os lados cliente e servidor. No lado cliente, os usuários acessam via internet ou intranet, por meio de um navegador, os módulos e o Sistema de Gerenciamento de Banco de Dados (SGBD) hospedados no servidor. Foi utilizado o SGBD MySQL e o servidor *web* Apache, ambos disponíveis gratuitamente no *site* dos seus desenvolvedores. A interface dos módulos foi desenvolvido em HTML para acesso via navegador. A programação da lógica dos módulos executada no servidor e o acesso à base de dados foi realizada com a linguagem PHP e Javascript para tarefas executadas no cliente.

O módulo Portal é a interface inicial do sistema, responsável pela validação de usuário e interface dos módulos de Supervisão e Análise.

Diante da necessidade de categorias mais específicas para IC, detectada após o primeiro ciclo do método *Design Science Research*, o Módulo de Supervisão foi desenvolvido para que os especialistas possam classificar notícias coletadas manualmente da *web*, verificar e validar as notícias coletadas e classificadas automaticamente pelo sistema. Para isso, inclui uma interface para os atributos e valores definidos na etapa de entendimento do problema: Tabelas 5, 6, 7 e 8, como mostra Figura 20.

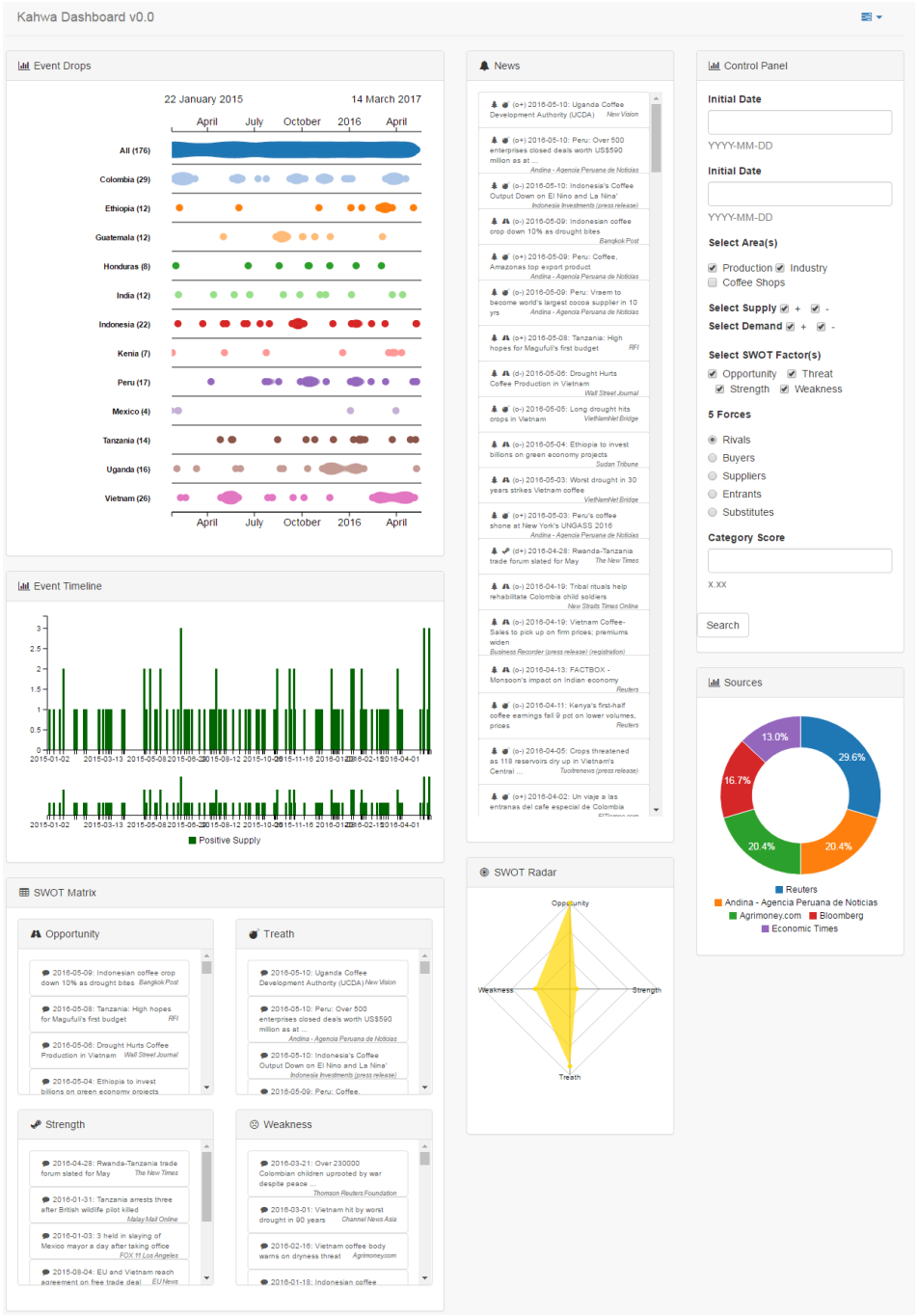
Figura 20 - Módulo de Supervisão.



O módulo de análise é um painel constituído de gráficos para visualização dos elementos propostos para análise. Para a construção dos gráficos, foram utilizados os recursos da biblioteca para Javascript D3js – https://d3js.org/.

A Figura 21 apresenta uma das visualizações possíveis dos dados para análise, nela é possível selecionar o período e os atributos das notícias coletadas pelo artefato e visualizar a ocorrência de notícias sobre os rivais no período selecionado, curva de oferta e demanda, gráfico de radar para SWOT, principais fontes das notícias e as notícias organizadas de acordo com os parâmetros selecionados para consulta.

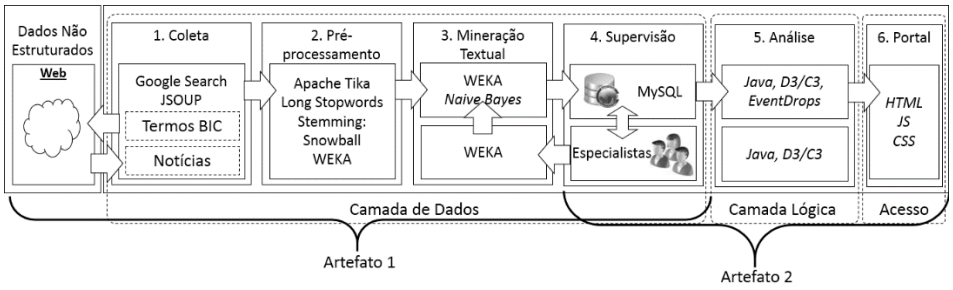
Figura 21 - Interface do Protótipo para análise de notícias.



4.4 Avaliação

Dois artefatos foram desenvolvidos como resultado de dois ciclos do método *Design Science Research*. O primeiro artefato teve como objetivo os módulos de coleta, pré-processamento e classificação de acordo com a cadeia produtiva, Figura 22, para que as notícias sejam organizadas nas dimensões analisadas pelos especialistas.

Figura 22 - divisão dos módulos em artefatos.



O segundo artefato teve como objetivo os módulos de Supervisão e Análise para incluir a perspectiva de Inteligência Competitiva.

4.4.1 Parâmetros para avaliação

Os resultados suscitam a avaliação de três aspectos: Tecnológico – desempenho da coleta e classificação, Análise – atendimento aos requisitos de inteligência e Qualidade do *Software* – dinamismo, flexibilidade, interoperabilidade e interface amigável entre outros. Entretanto, como é um sistema em desenvolvimento, os artefatos foram desenvolvidos como protótipos, e este aspecto será avaliado em trabalho futuro. O primeiro artefato foi avaliado quanto ao aspecto tecnológico o segundo foi avaliado quanto aos três aspectos.

4.4.1.1 Avaliação de aspectos tecnológicos

Para avaliar o desempenho dos módulos de coleta e classificação é necessário uma comparação entre a classificação automática e a realizada pelos especialistas para um mesmo conjunto de notícias. A Tabela 17 apresenta os resultados possíveis na comparação entre classificador e especialista para atribuir uma categoria C_i a uma notícia.

Tabela 17 - Comparação entre Classificador e Especialista para uma Categoria $C_{i.}$

Categoria 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 Sebastiani (2002).

Na Tabela 17, TP_i significa que o classificador acertou a classificação do especialista na categoria C_i, ou seja é um verdadeiro positivo para a categoria, enquanto FP_i é um falso positivo, indica que o classificador errou a categoria atribuída pelo especialista.

Conforme Sebastiani (2002), 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 ou Sensibilidade (ρ) propostas pelo autor e descritas para classificação textual por (BAEZA-YATES; RIBEIRO-NETO, 2013):

$$\pi i = \frac{TPi}{TPi + FPi}$$

e

$$\rho i = \frac{TPi}{TPi + FNi}$$

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

Precisão é a fração de todos os documentos atribuídos à classe C_i pelo classificador que realmente pertencem a classe C_i (de acordo com o conjunto de testes) e Revocação é a fração de todos os documentos que pertencem à classe C_i (de acordo com o conjunto de teste) que foram corretamente atribuídos à classe C_i pelo classificador (BAEZA-YATES; RIBEIRO-NETO, 2013). Assim, Precisão está relacionada à exatidão (qualidade) e revocação indica completude (quantidade).

E para obter estimativas de π e ρ as métricas:

$$\pi^{\mu} = \frac{\sum_{i=1}^{|C|} TPi}{\sum_{i=1}^{|C|}(TPi + FPi)}$$

e

$$\rho^{\mu} = \frac{\sum_{i=1}^{|C|} TPi}{\sum_{i=1}^{|C|}(TPi + FNi)}$$

Em que μ indica média.

Também são consideradas a Acurácia (Acc):

$$Acci = \frac{TPi + TNi}{TPi + TNi + FPi + FNi}$$

A fórmula *Acci* representa a quantidade de notícias classificadas corretamente em relação ao conjunto total de notícias classificadas. E a Medida F – a média harmônica entre Precisão e Revocação:

$$F = \frac{2 * \pi i * \rho i}{\pi i + \rho i}$$

As métricas de Precisão, Revocação e Medida F foram adotas para avaliar quantitativamente os artefatos pela perspectiva tecnológica, porém não são suficientes para demonstrar valor adquirido no processo de análise para IC. Por isso, foram definidos parâmetros qualitativos, descritos na próxima seção.

4.4.1.2 Avaliação de requisitos de inteligência

Em revisão sistemática, Cruz et al. (2015) apontam para a necessidade de estudos sobre indicadores de desempenho organizacional antes e após o processo de IC para ponderar sua efetividade. Bouthillier e Shearer (2003) propõem um modelo para avaliação de *software* para IC, entretanto, conforme Dai (2013), há as características de qualidade de *software* não específico para IC, assim não se encontrou consenso sobre critérios para a avaliação de IC bem estabelecidos.

Além disso, Teo e Choo (2001) apresentam evidência empírica de que a qualidade da informação para IC é positivamente relacionada com o impacto organizacional. Assim, foi proposto um questionário aplicado aos especialistas do BIC, elaborado com base nos critérios de Identificação de Necessidades e Aquisição de Informação Competitiva propostos por Bouthillier e Shearer (2003) para avaliar o artefato pela perspectiva de apoio e a obtenção de Inteligência Competitiva pelo estudo dos requisitos de inteligência definidos para a cafeicultura. As questões são apresentadas na Tabela 18.

Tabela 18 – Questionário para Avaliação.

Critérios	Perguntas
Identificação de Necessidades	As fontes das notícias são relevantes e frequentemente consultadas para análise? As notícias apresentadas são relevantes para o requisito de IC estudado?
Aquisição de Informação Competitiva	O sistema contribui para identificar oportunidades, ameaças, forças e fraquezas?
	O sistema contribui para identificar evidencias que impactam oferta e demanda?
	Pela análise das notícias é possível acrescentar informação competitiva não capturada sem o sistema?
	O artefato contribui para o requisito de IC?

Foram utilizadas questões abertas para que seja possível capturar no texto do especialista novas contribuições para o sistema.

4.4.1.3 Avaliação estatística

Para avaliar a relação entre as notícias sobre preço foi realizada uma análise descritiva com a ocorrência de notícias mensais sobre Produção e Indústria e sua correlação com a série de cotações do Café na Bolsa de Nova York (Café NY). Em outra abordagem foi utilizado o modelo ARIMA (BOX et al., 2015) para verificar a existência de alguma influência significativa de séries temporais formadas pela ocorrência de notícias classificadas para oferta e demanda sobre a série de volatilidade do preço na bolsa de Nova York.

Uma série temporal é um conjunto de informações ordenadas no tempo e representa uma possível realização de um processo estocástico. Existem modelos apropriados para séries financeiras que apresentam a variância evoluindo no tempo, neste caso, o modelo frequentemente utilizado na modelagem da variabilidade dos dados é a classe GARCH, como em Martins (2015), que consiste na generalização dos modelos ARCH (modelos

autorregressivos com heteroscedasticidade condicional) (BOLLERSLEV, 2008; ENGLE, 1982). Nestes modelos, é usual utilizar a série dos retornos, sendo o retorno definido como R= $\frac{P_t - P_{t-1}}{P_{t-1}}$, *em que* P_t é o preço do ativo no instante *t*.

Como o objetivo é verificar a influência de séries geradas pelas notícias em séries de preço e volatilidade, optou-se pelo modelo ARIMA, pela possibilidade de incluir as séries como covariáveis no modelo de forma intuitiva, sendo assim chamados de ARIMAX. Já nos modelos GARCH, a modelagem com covariáveis para relacionar as séries não é direta.

A volatilidade foi calculada e sua média modelada a partir do modelo ARIMAX, considerando covariáveis de interesse.

O modelo ARIMA, que consiste em ajustar modelos autorregressivos integrados de médias móveis, é representado por ARIMA (p, d, q), em que "p" refere-se à ordem de autorregressão, "d" a ordem de integração e "q" a ordem de média móvel. As ordens dos modelos ARIMA foram determinadas a partir da comparação de modelos com diferentes ordens, comparados pelo critério Akaike Information Criterion (AIC), sendo que as melhores ordens para os modelos ARIMA foram as que resultaram em menor AIC.

As medidas de acurácia para avaliar a capacidade preditiva foram considerando y_i o valor observado, $\hat{y}_i$ o valor predito pelo modelo, $\bar{y}$ a média dos valores observados e n o tamanho da amostra:

- A raiz do erro quadrático médio (RMSE) representa o desvio padrão amostral das diferenças entre os valores preditos e os observados, logo, quanto menor o RMSE melhor será o ajuste. O RMSE é calculado como:

  $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$.

- O desvio absoluto da média (MAD) representa o desvio dos valores ajustados em relação à média, logo, quanto menor o valor do MAD melhor será o ajuste. O MAD é calculado como:
  $MAD = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$.

- A média simétrica percentual absoluta do erro (SMAPE) expressa a acurácia do erro em relação à média. O SMAPE é dado em porcentagem e é calculado como:
  $SMAPE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n\bar{y}} \; x \; 100$.

Para comparar a qualidade da previsão (não atualizada) dos modelos sem covariáveis e com a previsão dos modelos com covariáveis, foi utilizado o teste de Wilcoxon Pareado (HOLLANDER; WOLFE; CHICKEN, 2013), sendo comparados os resíduos das previsões, ou seja, o valor predito pelo modelo subtraído do valor real. Com essa comparação busca-se verificar se o modelo sem covariáveis apresentou maiores resíduos que o modelo com covariáveis. Caso isto ocorresse, poderia concluir-se que o modelo com covariáveis apresentou um melhor poder preditivo.

Para verificar as suposições de que os resíduos do modelo ARIMA seguem uma distribuição normal e são independentes, foram utilizados o teste de normalidade de Jarque-Bera (JARQUE; BERA, 1980) e o teste de Ljung-Box (LJUNG; BOX, 1978).

O *software* utilizado nas análises foi o R (versão 3.1.3).

A avaliação dos classificadores, dos requisitos de inteligência e a influência das notícias no preço, de acordo com os critérios apresentados é descrita nas seções posteriores de acordo com cada módulo do artefato a que se referem.

4.4.2 Coleta, pré-processamento, supervisão e classificação (Artefato 1)

Para o primeiro artefato, os testes foram realizados inicialmente com 3.028 notícias em inglês já coletadas e classificadas pelos especialistas no período de 2011 a 2015, em seguida com notícias coletadas pelo sistema. A distribuição de notícias coletadas pelos especialistas é apresentada na Tabela 19.

Tabela 19 - Notícias Coletadas pelos Especialistas por Categoria.

Categoria	**Número de Notícias**
Bebidas	416
Cafeterias	1156
Indústria	752
Produção	704

4.4.2.1 Testes com a base de dados do BIC

Para selecionar um classificador, foram avaliadas as versões codificadas na ferramenta WEKA com parâmetros padrão, a saber: Bayesianos – Naïve Bayes (JOHN; LANGLEY, 1995), Naïve Bayes Multinomial (MCCALLUM; NIGAM, 1998), Complement Naïve Bayes (RENNIE et al., 2003), Discriminative Multinomial Naïve Bayes (SU et al., 2008); Árvore de Decisão – J48 (QUINLAN, 1993), Random Tree e Support Vector Machines – SMO (PLATT, 1998).

Na avaliação dos métodos, para treinamento e classificação, o banco de dados foi dividido em dois conjuntos de dados: $\Omega_t = \{d_1,...,d_{|\Omega t|}\}$, em que $|\Omega_t|$ representa o número de notícias compreendido entre duas datas para treinamento. E o conjunto: $\Omega_c = \{d_1,...,d_{|\Omega t|}\}$, em que $|\Omega_c|$ 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 = {*"Indústria", "Produção", "Cafeterias", "Bebidas"*} e os algoritmos classificadores representados pelo conjunto: $\phi = \{\phi_1,...,\phi_7\}$.

Com a descrição dos parâmetros anteriores, o procedimento foi para cada ϕ_n em ϕ, treine ϕ_n com Ω_t e para cada d_n em Ω_c classifique d_n em C com ϕ_n.

Para avaliar os resultados em situações diferentes, um experimento considerou o conjunto de treinamento Ω_t composto por 2.507 notícias em inglês coletadas entre 01/01/2011 e 31/12/2013 e para o conjunto de teste Ω_c, 387 notícias coletadas no período de 01/01/2014 a 31/12/2014, também em inglês.

Outro experimento verificou o comportamento dos algoritmos classificadores à medida que a base de treinamento aumentava. Assim, o conjunto de teste Ω_c foi composto pelas notícias extraídas no período de um mês – período usado para produzir um relatório do Bureau – de 01/12/2014 até 31/12/2014, totalizando um conjunto constante de 23 notícias em inglês, e o conjunto de treinamento foi escalonado retrocedendo mensalmente de novembro de 2014 a janeiro de 2012. O primeiro teste com um conjunto composto por 24 notícias de 01/11/2014 a 30/11/2012, o segundo no período de 01/10/2014 a 30/11/2014 com 81 notícias e assim sucessivamente até um conjunto com 2.056 notícias coletadas de 01/01/2012 até 30/11/2014 para treinamento.

Os resultados publicados por Lima Júnior, Castro Júnior e Zambalde (2015) mostram um desempenho satisfatório de *Naive Bayes*, portanto, este método foi utilizado no desenvolvimento do artefato para classificar as notícias coletadas da *web*, conforme descrito a seguir.

4.4.2.2 Teste com dados coletados da *web*

Neste teste, o módulo de coleta foi acionado para pesquisar termos montados pela concatenação da palavra *coffee* com cada uma das 132 palavras definidas pelos especialistas. Cada termo foi pesquisado em intervalo mensal a

partir de janeiro de 2011 até dezembro de 2015. Cada requisição de pesquisa por termo retorna uma página com 10 resultados de acordo com o critério de relevância do motor de busca. Cada notícia é então pré-processada, classificada e armazenada no banco de dados pelos respectivos módulos.

A execução do módulo de coleta com os parâmetros descritos resultou em 40.396 notícias. Foram excluídas 5.313 notícias sem o texto, o que ocorre por necessidade de validação de usuário para acesso ou remoção da notícia da URL indicada pelo motor de busca. As 35.083 válidas são distribuídas conforme representação na Tabela 20.

Tabela 20 – Distribuição das Notícias Coletadas pelo Módulo de Coleta.

Categoria	**Número de Notícias**
Bebidas	1994
Cafeterias	23867
Indústria	5027
Produção	4195
Total	35083

Do total de 3.028 notícias coletadas pelos especialistas de janeiro de 2011 a dezembro de 2015, 13,14% (398) também foram recuperadas da *web* pelo módulo de coleta. Entretanto, em 45 o sistema não teve acesso ao texto, portanto, não classificou. O resultado da classificação das 353 notícias comuns é apresentado na Tabela 21.

Tabela 21 – Resultado da Classificação para Categorias da Cadeia Produtiva.

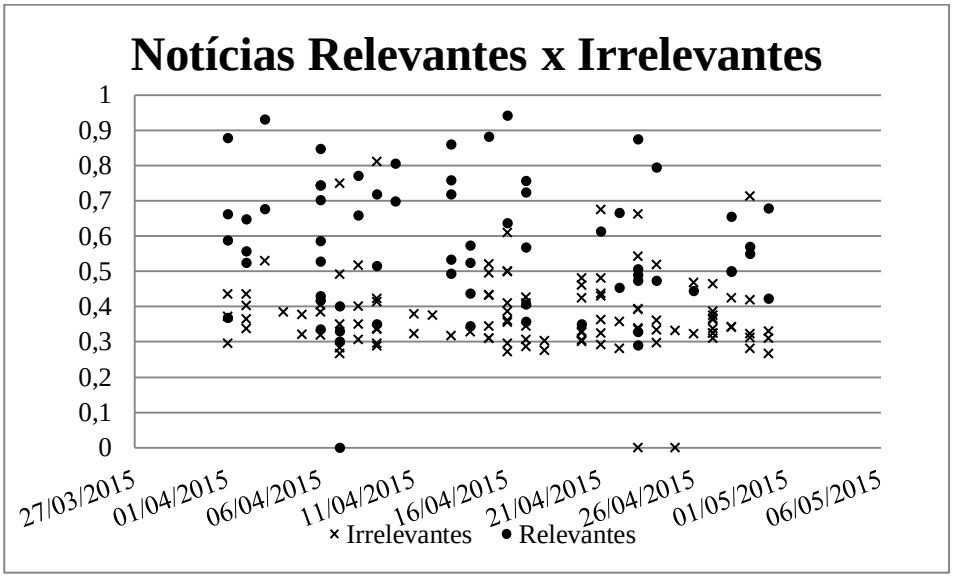
Categoria	Total	TP	FP	FN	Precisão	Revocação	Medida F
Bebidas	11	0	9	11	0	0	-
Cafeterias	123	107	41	16	0,72	0,86	0,78
Indústria	145	106	12	39	0,89	0,73	0,80
Produção	74	65	13	9	0,83	0,87	0,84

O classificador acertou 78% das notícias apesar da categoria Bebidas para a qual não obteve acertos e sim nove falsos positivos. Para as categorias Indústria e Produção, o resultado foi satisfatório. O número menor de amostras da categoria Bebidas para treinamento sugere que o resultado geral pode ser melhorado através de mais amostras desta categoria classificadas pelos especialistas no módulo de Supervisão.

Outro problema identificado é a falta de uma base de treinamento negativa. Como o algoritmo classificador sempre atribui uma das categorias para uma notícia, um número significativo de notícias irrelevantes é introduzido gerando ruído. Este problema sugere a criação de uma base de notícias irrelevantes para treinamento, o que pode ser realizado pelos especialistas após a coleta e classificação no módulo de Supervisão, tornando o processo semi supervisionado até que se tenha uma base adequada para que o classificador elimine notícias que não pertencem ao contexto.

Uma verificação da pontuação dada pelo algoritmo classificador a cada notícia para considerá-la estatisticamente próxima a determinada categoria, conforme fórmula (3) na Seção 5.2.3, mostra que este é um fator candidato para identificar notícias irrelevantes. A Figura 23 mostra 175 notícias coletadas pelo sistema em um mês aleatório, distribuídas por data, pontuação atribuída pelo classificador e relevância segundo avaliação do especialista.

Figura 23 - Distribuição de notícias relevantes e irrelevantes.



De 175 notícias coletadas pelo sistema e avaliadas por especialista, 108 foram consideradas irrelevantes para o contexto e 66 notícias foram consideradas relevantes. A maior concentração de notícias consideradas irrelevantes pelo especialista tem pontuação abaixo de 0,5 atribuída pelo algoritmo classificador, ou seja, probabilidade abaixo de 50% de ser da categoria classificada, enquanto as relevantes acima deste valor. A pontuação média das 108 notícias coletadas e consideradas irrelevantes pelos especialistas é de 0,38, enquanto a pontuação média das 66 relevantes é de 0,57. Este critério deve ser avaliado com mais rigor, pois mesmo notícias com pontuação máxima para determinada categoria podem ser falsos positivos. Entretanto, enquanto não há uma base de treinamento negativa, para diminuir o nível de ruído no resultado, serão consideradas notícias com pontuação máxima atribuída pelo algoritmo classificador, neste caso o número de notícias por categoria é apresentado na Tabela 22.

Tabela 22 – Notícias com Pontuação Máxima por Categoria.

Categoria	Nº de Notícias	Pontuação Máxima
Bebidas	1994	1217
Cafeterias	23867	10351
Indústria	5027	853
Produção	4195	1665
Total	35083	14086

O volume de notícias inviabiliza a validação manual de toda a base de dados, portanto, para avaliar a classificação automática, uma amostra de notícias coletadas e classificas pelo sistema foi apresentada aos especialistas. A amostra foi selecionada aleatoriamente com 100 notícias de cada categoria. O resultado foi significativo para as categorias Indústria e Produção com precisão de 89% e 69% respectivamente. Para as outras, o resultado foi insuficiente.

Portanto a construção do artefato mostra que é viável a utilização de Mineração Textual para auxiliar a coleta e organização dos dados para análise por setores da cadeia produtiva definidos pelos especialistas do BIC como Indústria e Produção. Aponta ainda a necessidade de uma base de treinamento negativa para identificar notícias irrelevantes, mais amostras da categoria Bebidas e balanceamento da base de treinamento.

4.4.3 Classificação para fatos relevantes, SWOT e oferta e demanda (artefato 2)

Pelo desenvolvimento do segundo artefato, foi possível testar os classificadores para atributos voltados à Inteligência Competitiva e funcionalidades do sistema para atender os requisitos de inteligência. Para isto, os especialistas classificaram um conjunto aleatório de notícias pelo módulo de Supervisão quanto aos critérios definidos para Fatos Relevantes (TABELA 8), SWOT (TABELA 6) e Oferta e Demanda (TABELA 5).

A classificação foi avaliada por validação cruzada com a base classificada pelos especialistas, por treinamento com notícias de 2011 a 2013 e classificação das notícias de 2014 e por treinamento com todas as notícias classificadas pelos especialistas e classificação das 35.083 notícias coletadas da *web* no Módulo de Coleta no desenvolvimento do Artefato 1, descrito na seção 7.1.

4.4.3.1 Classificador para Fatores

O primeiro teste para avaliação da classificação para Fatos Relevantes (TABELA 8) teve como objetivo analisar o desempenho do classificador Naive Bayes com o conjunto de teste formado por notícias classificadas pelos especialistas conforme Tabela 23.

Tabela 23 – Notícias Classificadas pelos Especialistas.

Categoria	**Número de Notícias**
Produção e Exportação	168
Eventos Naturais	41
Políticas	64
Indicação Geográfica	5
Expansão de Indústria	132
Expansão de Cafeterias	150
Consumo	84
Sustentabilidade	44
Empresas em Produtores	13
Pesquisa	11
Café Especial	15
Especulação	82

As categorias Indicação Geográfica, Empresas em Produtores, Pesquisa e Café Especial não foram incluídas no teste devido ao baixo número de

amostras. Com o objetivo de acrescentar ao modelo uma categoria para notícias que forem coletadas da *web* e não pertencem a nenhuma das classes, foi acrescentada a categoria Nenhuma, formada por 51 notícias avaliadas pelos especialistas, mas não classificadas em uma das categorias pré-definidas.

Outra característica é o desbalanceamento da base de treinamento o que interfere o desempenho da classificação e deve ser tratada por técnicas de balanceamento (GUIMARÃES, 2015). Assim antes do classificador, foi utilizado o método *Resample* da ferramenta WEKA para balanceamento da base de treinamento neste e nos demais testes.

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 distribuído por categorias é apresentado na Tabela 24.

Tabela 24 – Resultado da Classificação para Fatos.

Categoria	**TP Taxa**	**FP Taxa**	**Precisão**	**Revocação.**	**Medida F**
Produção e Exportação	0,786	0,095	0,64	0,786	0,706
Eventos Naturais	0,574	0	1	0,574	0,730
Políticas	0,733	0,025	0,698	0,733	0,715
Expansão da Indústria	0,839	0,040	0,810	0,839	0,824
Expansão de Cafeterias	0,856	0,035	0,861	0,856	0,859
Consumo	0,766	0,030	0,766	0,766	0,766
Sustentabilidade	0,860	0,009	0,860	0,860	0,860
Especulação	0,696	0,031	0,676	0,696	0,686
Nenhuma	0,489	0,003	0,786	0,489	0,639

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

De 816 notícias, 629 foram corretamente classificadas (77,08%). O resultado geral é satisfatório. As categorias Eventos Naturais e Nenhuma apresentaram baixos valores para Revocação.

No segundo teste, o conjunto de treinamento foi formado por 577 notícias de 01/01/2011 a 31/12/2013 e o conjunto de teste formado por 74 notícias de 01/01/2014 a 31/12/2014 classificadas pelos especialistas. A Tabela 25 apresenta a distribuição de notícias por categoria para treinamento e teste.

Tabela 25 - Distribuição de notícias por categoria.

Categoria	Treinamento (2011 a 2013)	Teste (2014)
1 Produção e Exportação	158	10
2 Eventos Naturais	39	2
3 Políticas	54	10
5 Expansão da Indústria	69	10
6 Expansão de Cafeterias	92	22
7 Consumo	52	13
8 Sustentabilidade	30	4
14 Especulação	80	2
Nenhuma	33	5
Total	577	74

As notícias do ano de 2014 foram classificadas pelo método *Naive Bayes* e comparadas com a classificação do especialista. O resultado distribuído por categorias é apresentado na Tabela 26.

Tabela 26 - Classificador *Naive Bayes* para Fatos relevantes.

Categorias	TP	FP	FN	Precisão	Revocação	Medida F
Produção e Exportação	2	6	8	0,250	0,200	0,22
Eventos Naturais	0	5	2	0	0	-
Políticas	5	7	5	0,417	0,500	0,455
Expansão da Indústria	2	0	8	1	0,200	0,333
Expansão de Cafeterias	21	11	1	0,656	0,955	0,778
Consumo	6	3	7	0,667	0,462	0,545
Especulação	1	2	1	0,333	0,500	0,400
Nenhuma	2	1	3	0,667	0,400	0,500

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

De 74 notícias classificadas, o método acertou 41 (55,41%). Para Expansão de Cafeterias e Consumo o classificador obteve níveis razoáveis de Precisão e Revocação. Apesar da Precisão 1 para Expansão da Indústria, o número de falsos negativos resulta em uma revocação que a compromete (0,20).

O resultado mostra que a amostra de treinamento deve ser melhorada, mesmo considerando poucas amostras para teste nos casos de Eventos Naturais, Sustentabilidade, Especulação e Nenhuma.

4.4.3.2 Classificador para SWOT

O primeiro teste teve como objetivo analisar o desempenho do classificador com o conjunto formado por notícias classificadas pelos especialistas, conforme Tabela 27.

Tabela 27 – Notícias Classificadas pelos Especialistas.

Categoria	**Número de Notícias**
Oportunidade	273
Ameaça	218
Força	102
Fraqueza	36

Nesta etapa, 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. Os resultados são apresentados na Tabela 28.

Tabela 28 - Resultados da Validação Cruzada.

Categoria	**Taxa TP**	**Taxa FP**	**Precisão**	**Revocação**	**Media-F**
Oportunidade	0,672	0,143	0,702	0,672	0,687
Ameaça	0,744	0,124	0,704	0,744	0,723
Fraqueza	0,444	0,008	0,727	0,444	0,552
Força	0,505	0,07	0,519	0,505	0,512
Nenhuma	0,594	0,131	0,543	0,594	0,577

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

Do total de 821 notícias, 529 foram classificadas corretamente (64,43%). O desempenho geral é aceitável, entretanto, o resultado mostra que as categorias Oportunidade e Ameaça possuem melhor precisão e revocação enquanto as outras classes apresentam desempenho não satisfatório, comprometendo o resultado geral.

Nesta etapa, a base de treinamento foi composta por 651 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 em categorias conforme demonstrado na Tabela 29.

Tabela 29 - Distribuição de notícias para treino e teste.

Categorias	**Treino**	**Teste**
Oportunidade	198	35
Ameaça	195	18
Força	62	5
Fraqueza	28	4
Nenhuma	135	16

As notícias do ano de 2014 foram classificadas pelo método *Naive Bayes* e comparadas com a classificação do especialista em métricas apresentadas anteriormente nesta seção. O resultado distribuído por categorias é apresentado na Tabela 30.

Tabela 30 - Resultado da Classificação distribuído por categorias.

Categoria	**TP**	**FP**	**FN**	**Precisão**	**Revocação**	**Medida F**
Oportunidade	18	10	17	0,634	0,514	0,571
Ameaça	8	15	10	0,348	0,444	0,390
Força	0	4	4	0	0	-
Fraqueza	2	15	3	0,118	0,400	0,182
Nenhuma	1	5	15	0,167	0,063	0,091

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

Do total de 78 notícias, 29 foram classificadas corretamente (37,18%). O desempenho geral é insatisfatório, entretanto, o resultado mostra que a categoria Oportunidade tem precisão e revocação aceitáveis, enquanto as outras classes apresentam desempenho não satisfatório. É importante ressaltar que há uma diferença de desempenho em relação a categorias específicas e que o conjunto de teste possui apenas quatro notícias classificadas como Fraqueza e cinco como

Força, este desbalanceamento deve ser considerado para evitar distorção na avaliação do resultado.

Para classificar 35.083 notícias coletadas pelo sistema, a base de treinamento foi composta por 812 notícias de 01/01/2011 até 31/01/2015, conforme Tabela 31.

Tabela 31 - Distribuição da base de treinamento em Categorias.

Categorias	**Treino**
Oportunidade	273
Ameaça	218
Força	102
Fraqueza	36
Nenhuma	192

Para viabilizar a análise, foram consideradas as notícias relevantes para a produção de acordo com o classificador para categorias da cadeia produtiva, descrito por Lima Júnior et al. (2015), e o intervalo de dois meses – outubro e novembro de 2015, totalizando 95 notícias. A Tabela 32 apresenta uma amostra de notícias coletadas corretamente como Oportunidades e Ameaças para o setor de Produção no mês de novembro de 2015.

Tabela 32 - Títulos de notícias separadas pelos fatores SWOT.

Oportunidade	**Ameaça**
"Colombia coffee harvest hit by El Nino drought"	*"Vietnam's coffee exports to rebound to record high"*
"Farmers face 'ruin' as drought threatens Colombian coffee harvest"	*"Kenya aims to double coffee output by 2020 after long decline"*
"Unpredictable weather takes heavy toll on coffee farmers in Uganda"	*"How a New Coffee Crop Is Helping a Dominican Community Thrive"*
"Climate changes take heavy toll on Ugandan coffee farmers"	

O conjunto classificado pelo sistema foi comparado com o julgamento dos especialistas para as mesmas notícias. O resultado é apresentado na Tabela 33, dividido por meses para permitir uma avaliação em sincronia com a produção mensal de relatórios do BIC.

Tabela 33 - Resultado da Classificação distribuído por categorias.

Data	Cat.	Total	TP	FP	FN	Precisão	Revocação	Medida F
Out. 2015	O	17	8	7	9	0,533	0,470	0,500
	T	16	8	5	8	0,615	0,500	0,551
	S	0	0	4	0	-	-	-
	W	0	0	6	0	-	-	-
	N	12	0	8	11	0	0	-
Nov. 2015	O	18	10	11	8	0,476	0,555	0,512
	T	12	7	8	5	0,466	0,583	0,518
	S	1	0	3	1	0	0	-
	W	0	0	8	0	-	-	-
	N	19	0	3	19	0	0	-

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

A precisão, revocação e medida F para as categorias Oportunidade e Ameaça foram próximas as obtidas nos testes anteriores. A consulta com os parâmetros descritos nesta etapa retornaram apenas uma notícia sobre a categoria Força e nenhuma sobre Fraqueza nos meses considerados.

4.4.3.3 Classificador para Oferta e Demanda

O primeiro teste para avaliação da classificação de notícias quanto a evidências positivas ou negativas para oferta e demanda (TABELA 5) teve como objetivo analisar o desempenho do classificador Naive Bayes com o conjunto de teste formado pelas 861 notícias classificadas pelos especialistas conforme Tabela 34.

Tabela 34 – Notícias Classificadas pelos Especialistas para Oferta e Demanda.

Categoria	**Número de Notícias**
Oferta Positivo (aumento de oferta)	83
Oferta Negativo (diminuição de oferta)	73
Demanda Positivo (aumento de demanda)	149
Demanda Negativo (diminuição de demanda)	48
Nenhuma	200

O resultado da validação cruzada na ferramenta WEKA é apresentado na Tabela 35.

Tabela 35 – Validação Cruzada para Classificação de Oferta e Demanda.

Categoria	**Taxa TP**	**Taxa FP**	**Precisão**	**Revocação**	**Medida F**
Oferta Positivo	0,653	0,061	0,628	0,653	0,641
Oferta Negativo	0,893	0,127	0,754	0,893	0,817
Demanda Positivo	0,531	0,07	0,5	0,531	0,515
Demanda Negativo	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 categorias Oferta Negativo e Demanda Positivo possuem melhor precisão enquanto as outras classes apresentam desempenho não satisfatório comprometendo o resultado geral.

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 em categorias conforme Tabela 36.

Tabela 36 - Distribuição de notícias para treino e teste.

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

As notícias do ano de 2014 foram classificadas pelo método *Naive Bayes* e comparadas com a classificação do especialista em métricas apresentadas anteriormente nesta seção. O resultado distribuído por categorias é apresentado na Tabela 37.

Tabela 37 - Resultado da Classificação distribuído por categorias.

Categoria	TP	FP	FN	Precisão	Revocação	Medida F
Oferta Positivo	2	3	7	0,400	0,222	0,286
Demanda Positivo	21	10	7	0,677	0,750	0,712
Oferta Negativo	4	4	4	0,500	0,500	0,500
Demanda Negativo	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 categorias Demanda Negativo, Oferta Positivo e Nenhuma que apresenta baixo nível de Revocação.

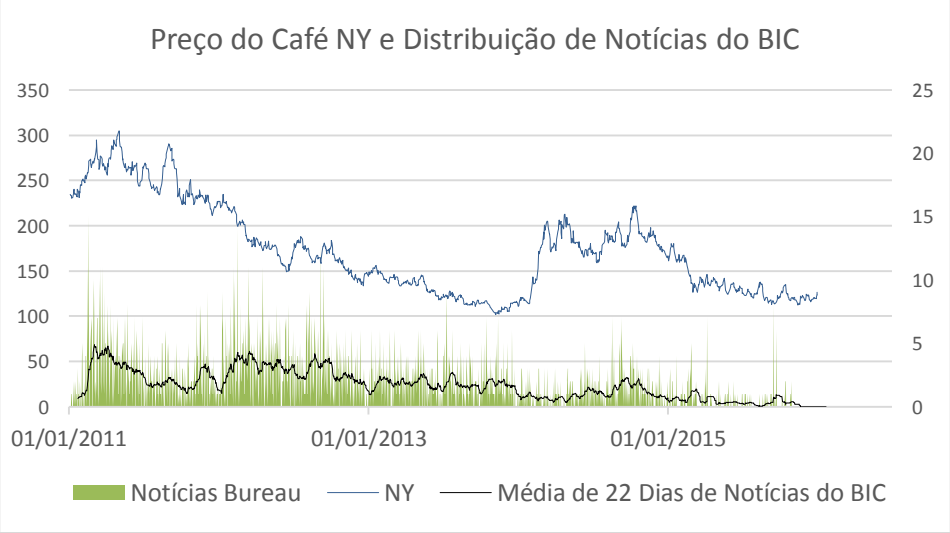
4.4.4 Avaliação preço e volatilidade

Com o objetivo de atender ao requisito de inteligência sobre a volatilidade do preço do café no mercado, a classificação das notícias positivas e negativas para oferta e demanda foi distribuída cronologicamente para obter uma

linha de eventos e permitir um estudo de séries temporais em relação ao preço do café e volatilidade do preço.

A Figura 24 apresenta a distribuição das notícias do Bureau no tempo – área cinza, a variação preço do café na BM&F – linha pontilhada, e a variação de preço do café na Bolsa de Nova York – linha preta grossa. Visualmente percebe-se uma relação entre o preço nas duas bolsas.

Figura 24 - Preço do Café NY x Notícias coletadas pelo BIC.

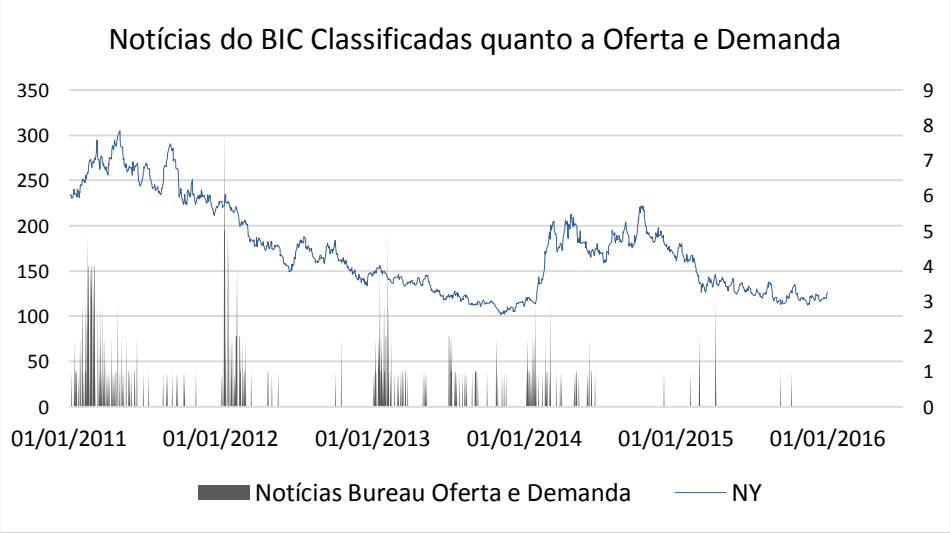


A linha preta mais fina é a média móvel de 22 períodos das notícias do Bureau, nela é possível observar períodos em que a ocorrência de notícias cai subitamente. Uma vez que a curva reflete o comportamento dos especialistas na coleta de notícias, estes períodos de queda que ocorreram na mesma época, janeiro de 2012, 2013 e 2014, são explicados por recessos, período em que o Bureau diminui o número de buscas na *web* e não por uma provável escassez de notícias na *web*, já que os fatos são constantes e as agências de notícias não

param. Uma análise utilizando esta base é distorcida por estas lacunas, o que reforça a necessidade do módulo de coleta automática de notícias.

Um conjunto de notícias foi aleatoriamente selecionado e classificado pelos especialistas quanto a impacto para oferta e demanda. A Figura 25 apresenta a distribuição de 353 notícias deste conjunto e o preço do café na Bolsa de Nova York.

Figura 25 - Café NY x notícias classificadas pelo BIC em oferta e demanda.



Estas notícias foram utilizadas para gerar o modelo de treinamento para o classificador de oferta e demanda, que classificou 35.083 notícias válidas selecionadas da *web* pelo módulo de coleta. Assim, foram geradas quatro séries temporais de ocorrências de notícias por classificação automática:

a) *Web* o+: Soma de ocorrências de notícias classificadas como positivas para oferta no dia;

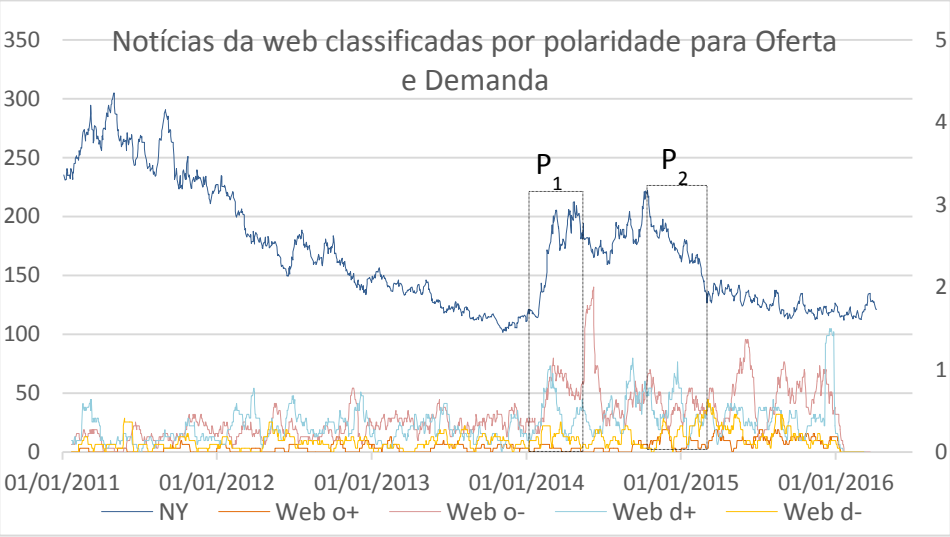
b) *Web* o-: Soma de ocorrências de notícias classificadas como negativas para oferta no dia;
c) *Web* d+: Soma de ocorrências de notícias classificadas como positivas para demanda no dia;
d) *Web* d-: Soma de ocorrências de notícias classificadas como negativas para demanda no dia.

A avaliação dos classificadores mostrou que as categorias Bebidas e Cafeterias têm níveis de Precisão e Revocação que precisam ser melhorados, portanto, as séries consideradas para o estudo incluíram as ocorrências de notícias classificadas automaticamente como Produção e Indústria.

A Figura 26 apresenta quatro séries formadas pela média móvel de 22 períodos de cada série gerada, intervalo que abrange um mês de pregão e análises do BIC. São elas:

a) Oferta Positivo: média de 22 períodos de *Web* o+;
b) Oferta Negativo: média de 22 períodos de *Web* o-;
c) Demanda Positivo: média de 22 períodos de *Web* d+;
d) Demanda Negativo: média de 22 períodos de *Web* d-;

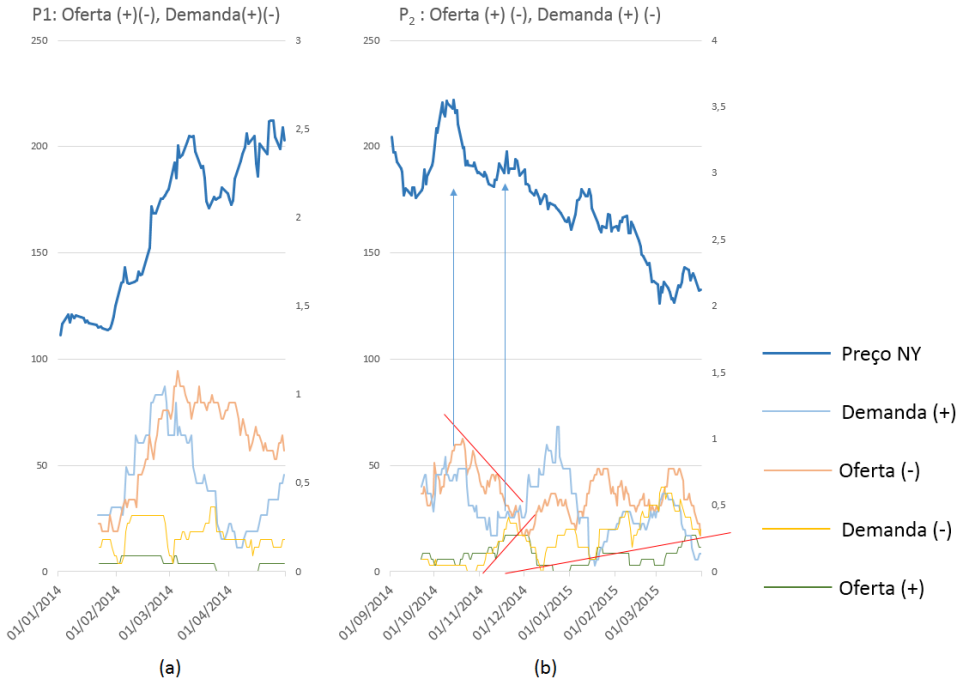
Figura 26 – Notícias Coletadas da Web e Classificadas quanto a Oferta e Demanda.



Para observar o comportamento das séries, dois momentos, marcados na Figura 26 como P_1 e P_2, foram selecionados pelo aumento da volatilidade, em P_1 com alta no preço após longa tendência de baixa e em P_2 queda após não ultrapassar a última região de preço atingida.

A Figura 27 tem as duas regiões como foco.

Figura 27 – Regiões do Gráfico de Notícias Coletadas da Web.



A Figura 27 (a) mostra o período P_1. As séries que convergem com o movimento do preço neste período são Demanda (+) e Oferta (-), que representam respectivamente a média de ocorrência de notícias nos últimos 22 dias que sugerem aumento de demanda (*Web* d+) e diminuição de oferta (*Web* o-). O mesmo ocorre em P_2, as curvas de Demanda (+) e Oferta (-) convergem com a queda de preço, enquanto a Demanda (-) (média de 22 períodos de *Web* d-) inicia uma tendência de alta no período de queda do preço.

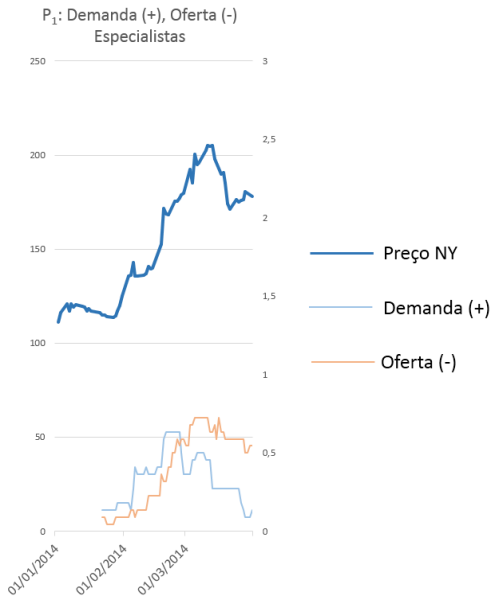
A Tabela 38 apresenta uma amostra de notícias que geraram as séries de Demanda (+) e Oferta (-) respectivamente 45 e 61 notícias, no período P_1, de 01/01/2014 a 31/03/2014.

Tabela 38 – Amostra de Notícias para Demanda Positiva e Oferta Negativa.

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

A classificação foi submetida à avaliação dos especialistas como teste para o classificador de Oferta e Demanda no caso de notícias coletadas da *web*. A avaliação apontou 25 notícias corretamente classificadas entre as 45 como Demanda Positiva (55,55%) e 35 corretamente classificadas entre as 61 como Oferta Negativa (57,37%). Resultado satisfatório considerando 5 categorias para o classificador. A Figura 28 apresenta a região P1 com a distribuição das notícias consideradas corretamente classificadas, sendo possível observar que o comportamento permanece.

Figura 28 – Região P1 do Gráfico de Notícias para Oferta Negativa e Demanda Positiva.



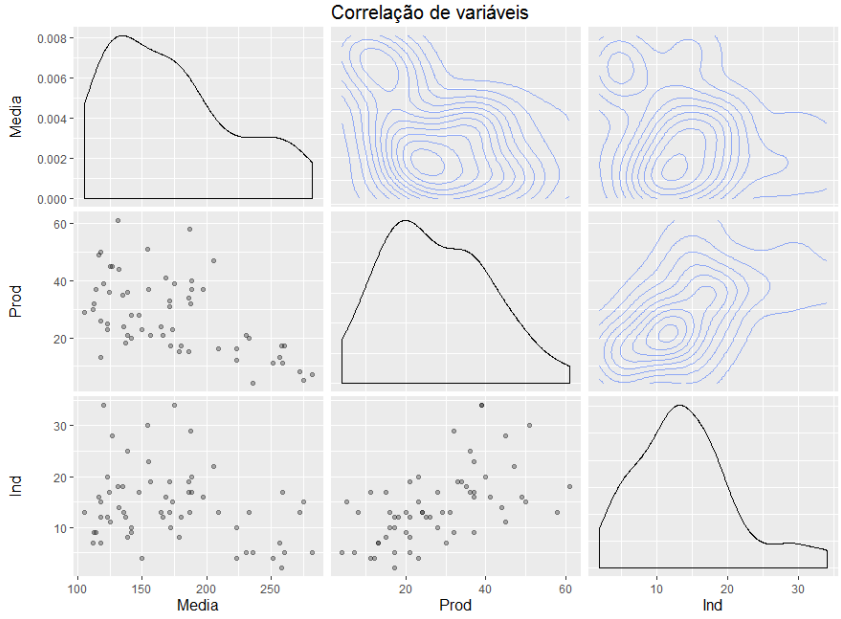
O cenário delineado pelas séries é coerente com os fundamentos do mercado e aconteceu em uma das lacunas de notícias do BIC (P_1 mostrado na Figura 28), entretanto é uma visão especulativa que se aproxima de divergências da Análise Gráfica, mas remete a um tratamento estatístico, apresentado na próxima seção.

4.4.4.1 Relação entre notícias e preço

Para verificar a correlação entre notícias e preço, foi realizada uma análise descritiva com os dados mensais de 2011 a 2015 incluindo média de preço mensal, número de notícias no mês classificadas como Produção, com pontuação máxima pelo classificador, e número de notícias no mês classificadas como Produção com pontuação máxima pelo classificador. Estes dados são apresentados no Anexo D. O resultado apresentado na Figura 29, sugere uma

correlação inversa entre notícias mensais da produção e média de preços mensais.

Figura 29 – Correlação de Variáveis.



O teste de Pearson para média mensal de preço e notícias mensais da produção resulta em coeficiente de -0,558 com p-valor de 3,50e-06 e intervalo de confiança -0,711 e -0,354, confirmando a correlação observada no gráfico. No caso de média mensal e notícias da Indústria, o teste de Pearson apresenta resultado menos significativo de -0,28 com p-valor 0,029 com intervalo de confiança de -0,50 e -0,03.

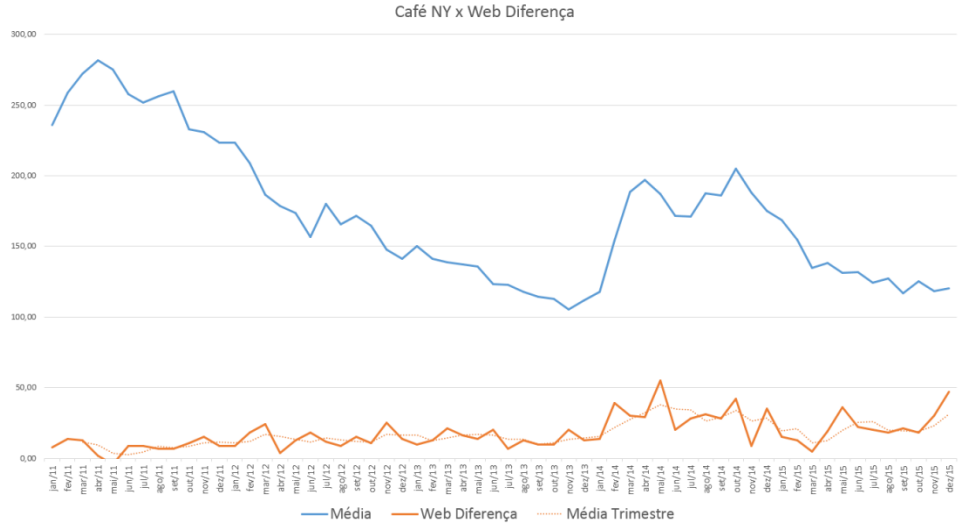
Acrescentando as notícias mensais da Produção classificadas também como positivas para Oferta, o resultado do teste de Pearson resulta em -0,45 com p-valor de 0,00030 o que é coerente com o fundamento de preços menores com aumento de oferta.

A verificação não implica causalidade, mas os dados revelam uma correlação inversa significativa, em que um cenário com mais notícias de Produção, os preços estão em patamares mais baixos.

Com o objetivo de verificar a existência de influência significativa entre a ocorrência de notícias, coletadas pelo artefato, sobre o preço do café de acordo com conceitos de oferta e demanda que exercem influência para alta de preço (oferta negativa e demanda positiva) e queda de preço (oferta positiva e demanda negativa) foi criada a série *Web* diferença, calculada a partir dos valores mensais das séries de ocorrências classificadas automaticamente:

Web diferença Mensal = [(*Web* o-) + (*Web* d+)] – [(*Web* o+) + (*Web* d-)]

Figura 30 - Café NY mensal x ocorrência de notícias na *web*.



A Figura 30 apresenta a variação do preço médio mensal do café na bolsa de NY e a série *Web* Diferença Mensal (WDM). É possível observar uma variação maior do preço no primeiro trimestre de 2014 e no mesmo período na série *Web* Diferença Mensal.

Assim como primeiro passo, a análise de pontos extremos (*outliers*), na série *Web* Diferença Mensal com modelo ARIMA (0,1,1) sem constante, mostrou os pontos na Tabela 39.

Tabela 39 – *Outliers.*

	WDM	**Ajustado**	**Resíduo**
2014/02	39.00	13.89	25.11*
2014/05	55.00	25.10	29.90*

Nota: * assinala um resíduo que excede 2,5 vezes o erro padrão.

O impacto dos *outliers* em WDM é mostrado na Tabela 40.

Tabela 40 – Impacto dos *Outliers.*

	coeficiente	**Erro padrão**	**z**	**p-valor**
Theta_1	-0.761376	0.0971116	-7.840	4.5e-015***
i1_WDM(2014/02)	18.6758	7.82434	2.387	0.0170**
i2_WDM(2014/05)	31.7330	7.82435	4.056	5.00e-05***

Como segundo passo, para testar se os *outliers* em WDM têm capacidade preditiva do comportamento dos preços, foi utilizado um modelo AR(1) do tipo Prais-Winsten (Prais e Winsten, 1954) para o primeiro *outlier*, em fevereiro de 2014 (2014/02) usando as observações de janeiro de 2011 a dezembro de 2015 (T = 60) e como variável independente, o log da Média de preços mensais resultando em rho = 0.980193 e Tabela 41.

Tabela 41 – Resultado do Modelo.

	Coeficiente	**Erro padrão**	**razão-t**	**p-valor**
Const	5.03650	0.273997	18.38	2.17e-025***
i1	0.199042	0.0796775	2.498	0.0154**
Wdm	-0.00281142	0.00153569	-1.831	0.0725*
i3	0.00334865	0.00178640	1.875	0.0661*

Na Tabela 41, i1 é a varável de intervenção (*dummy*), assume o valor 1 para o intervalo de fevereiro de 2014 (2014/02) até dezembro de 2015 (2015/12) e zero, caso contrário. Indica que nesse período ocorreu uma quebra na série (mudança de intercepto). A série de preço passou a correr em um patamar 0,199 mais elevado.

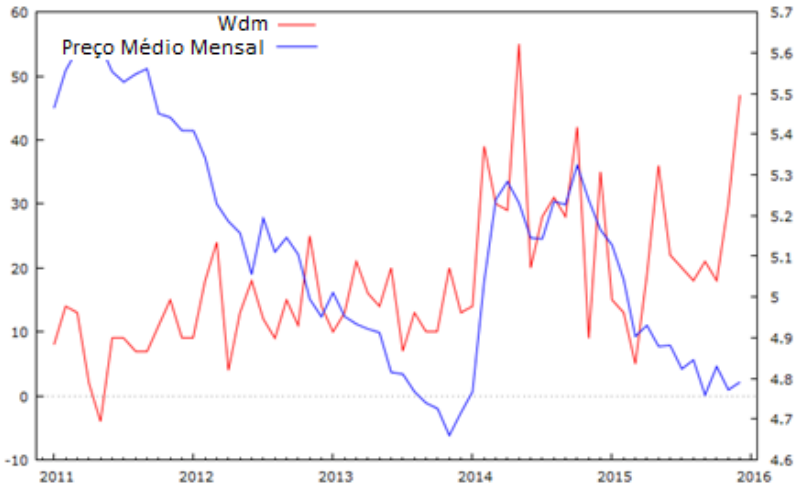
Wdm indica que a série WDM contribuiu para a explicação do preço, ainda que de forma pouco significativa (0,05<p-valor<0,10) no período anterior a 2014/02, com um coeficiente de −0.00281142. A cada ponto adicional de WDM o preço tende a cair 0,28%.

A variável de interação i3 = wdm*i1 (interação entre o wdm e a intervenção i1) indica que wdm contribuiu para a explicação dos preços no período pós 2014/02, com um coeficiente de −0.00281142+0.00334865 = 0,000537. A cada ponto adicional de WDM, o preço tende a subir 0,05%.

O teste do *outlier* em maio de 2014 (2014:05) não foi significativo.

A Figura 31 apresenta a relação entre Wdm e o logaritmo da média mensal de preços do café na bolsa de Nova York.

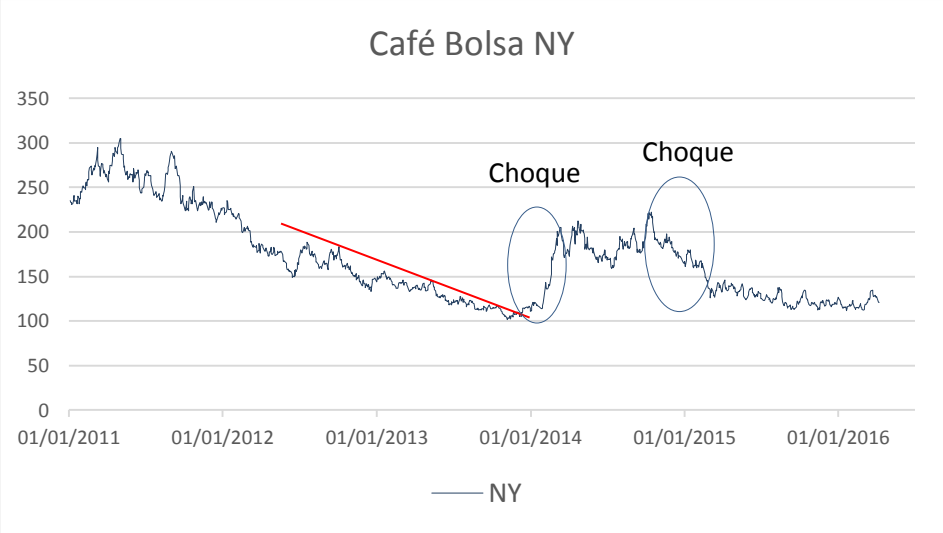
Figura 31 - Relação entre wdm e ln(preço médio mensal).



4.4.4.2 Análise da série volatilidade NY

Conforme Figura 32, há períodos de choques menores na série de preço com equilíbrio alcançado em prazos mais curtos e choques maiores com equilíbrio em prazo mais longo, cujo ajuste, segundo Nunes, Saes e Brando (2004), é plausível uma alternância de trechos como passeio aleatório em pequenos choques (próximo ao equilíbrio de longo prazo) e séries auto regressivas no ajuste a choques maiores (em direção ao equilíbrio de longo prazo).

Figura 32 - Preço do Café na Bolsa de Nova York.



Pode-se observar períodos com maior amplitude das diferenças entre os dias (regiões de Choque na FIGURA 32), e períodos com menor diferença. Se os preços do café estão em passeio aleatório as diferenças do preço atual (p_t) e preço do dia anterior (p_t-1) representam uma medida do tamanho de ε, e da volatilidade.

Assim, para verificar a relação de notícias com a volatilidade do preço foi incluída a série:

Volatilidade NY: volatilidade da série Café NY. Calculado conforme indicador apresentado em Nunes, Saes e Brando (2004) para um mês (22 pregões) como a média móvel da raiz quadrada das variações percentuais entre dias, elevadas ao quadrado.

A série Café NY foi selecionada pela maior liquidez e volume negociado.

Para comparar o resultado do artefato com uma abordagem por Análise de Sentimento foi utilizada uma série gerada pela aplicação do método Combined (GONÇALVES et al., 2013) a trechos de notícias coletadas pelos especialistas de 2011 a outubro de 2015.

Sentimento Combined: Soma de ocorrências diárias de trechos de notícias classificadas quanto a sua polaridade: positivas, negativas ou neutras.

A influência das séries *Web* o+, *Web* o-, *Web* d+, *Web* d- e Sentimento sobre a Volatilidade NY foi avaliada em quatro cenários distintos. No primeiro cenário, a informação das covariáveis *Web* diferença, *Web* o+, *Web* o-, *Web* d+, *Web* d- e Sentimento foi inserida como o valor diário da série, no segundo cenário foram considerados valores das covariáveis como a média dos últimos sete dias e no terceiro cenário a informação das covariáveis como a média dos últimos 22 dias.

No quarto cenário, foi realizada uma análise mensal, sendo considerados os valores médios mensais da Volatilidade NY e os valores das covariáveis como a média do mês anterior.

Com o intuito de comparar os modelos com e sem as covariáveis *Web* diferença, *Web* o+, *Web* o-, *Web* d+, *Web* d- e Sentimento os dados foram divididos em duas partes. Para os três primeiros cenários, a divisão foi:

a) Período considerado para a estimação dos modelos: de janeiro de 2011 a novembro de 2015;
b) Período considerado para realizar as previsões: 15 primeiros dias úteis de dezembro de 2015.

Para estimação no cenário mensal, foi considerado o período de 2011 a julho de 2015 e para previsão os cinco próximos meses: de agosto de 2015 a dezembro de 2015.

A capacidade preditiva foi avaliada em duas situações: planejamento – em que as previsões ocorreram sem as devidas atualizações (sem resultados futuros para atualizar); e acompanhamento diário/mensal – em que as previsões são atualizadas a cada fechamento do dia ou mês. Para ambas as situações foram utilizadas as medidas de acurácia descritas no início da seção para os próximos 15 dias ou próximos cinco meses.

Calculou-se as médias dos últimos sete e 22 dias das covariáveis *Web* o+, *Web* o-, *Web* d+, *Web* d-, *Web* diferença e Sentimento. Sábados, domingos e feriados, no Brasil e Nova York, sem informações foram excluídos. Os dados faltantes em Café NY (64 *missing*) para calcular Volatilidade NY foram estimados pelo método Expectation Maximization (EM) (DEMPSTER et al., 1977), procedimento necessário pois a modelagem estatística com modelos como ARIMA, que utilizam informações passadas é prejudicada por dados faltantes.

Na construção do banco mensal, após realizada a entrada dos dados faltantes, calculou-se os valores médios mensais de todas as séries,

transformando a série diária em mensal. Posteriormente, a transformação diária para mensal, as covariáveis *Web* diferença, *Web* o+, *Web* o-, *Web* d+, *Web* d- e Sentimento foram defasadas um passo atrás.

A Tabela 42 apresenta a análise descritiva das séries em estudo. Dessa forma, pode-se destacar que:

a) A série Volatilidade NY apresentou valor médio igual a 6,95, sendo o valor mínimo de 2,32 e o máximo 16,03;
b) A série Sentimento Combined apresentou valor médio de 1,16 com desvio padrão (D.P.) igual a 1,77;
c) A série *Web* diferença apresentou valor médio igual a 0,72, sendo o valor mínimo -5,00 e o máximo 19,00;
d) As séries *Web* o+ P1, *Web* o- PI, *Web* d+ PI e *Web* d- PI apresentaram valor médio de 0,07; 0,51; 0,46 e 0,18 respectivamente.

Tabela 42 - Análise descritiva das séries de interesse.

Séries	**Média**	**D.P.**	**Mín.**	**1Q**	**2Q**	**3Q**	**Máx.**
Volatilidade NY	6,95	2,32	2,63	5,40	6,51	8,09	16,03
Sentimento Combined	1,16	1,77	-3,00	0,00	1,00	2,00	12,00
Web diferença	0,72	1,33	-5,00	0,00	0,00	1,00	19,00
Web o+ PI	0,07	0,28	0,00	0,00	0,00	0,00	3,00
Web o- PI	0,51	0,79	0,00	0,00	0,00	1,00	6,00
Web d+ PI	0,46	0,96	0,00	0,00	0,00	1,00	18,00
Web d- PI	0,18	0,48	0,00	0,00	0,00	0,00	5,00

Fonte: Dados da pesquisa

A série Volatilidade NY é apresentada no Gráfico 1 e os Gráficos 2, 3 e 4 mostram esta série com as séries *Web* o+, *Web* o-, *Web* d+, *Web* d- e

Sentimento. Não há nenhuma relação clara entre a série Volatilidade NY e as séries das covariáveis.

Gráfico 1 - Série Volatilidade NY.

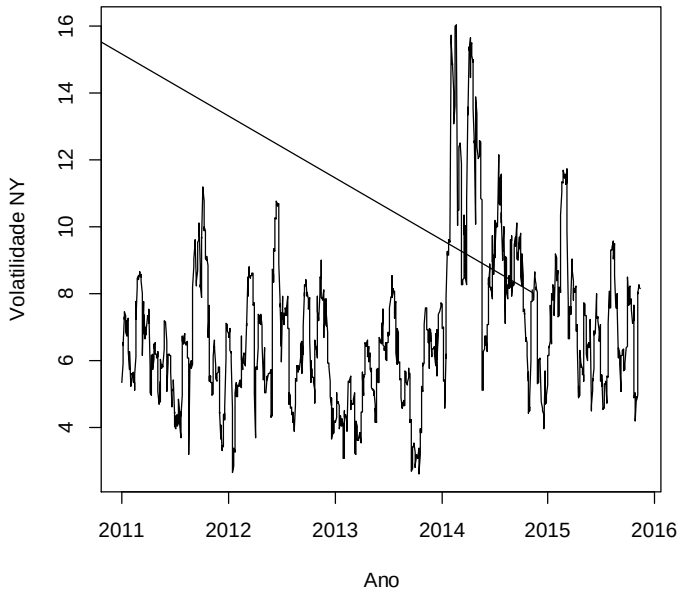


Gráfico 2 - Série Volatilidade NY e séries Web o+ e Web o-.

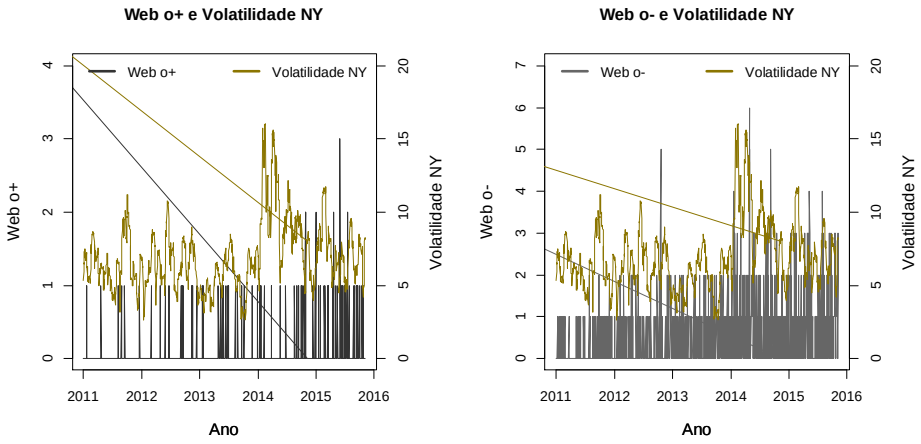


Gráfico 3 - Série Volatilidade NY e séries Web d+ e Web d-.

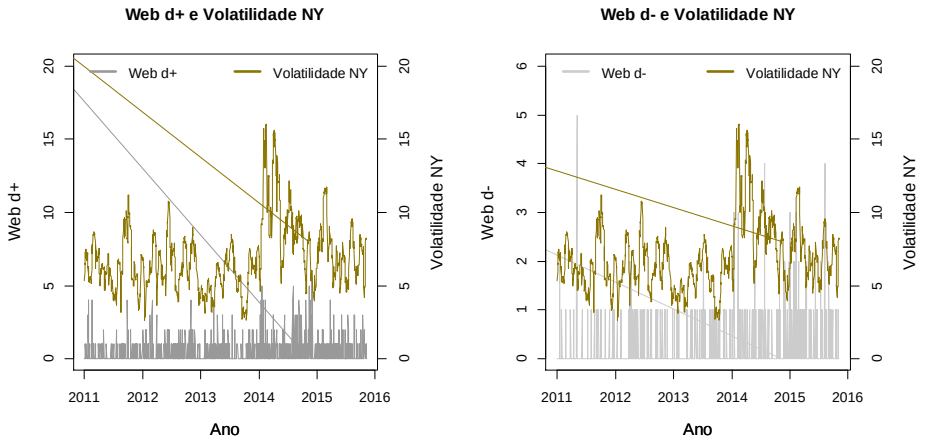
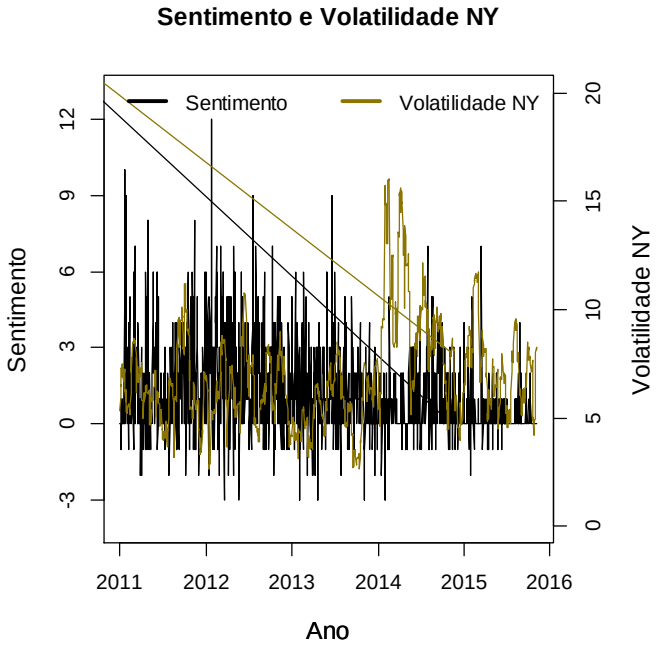
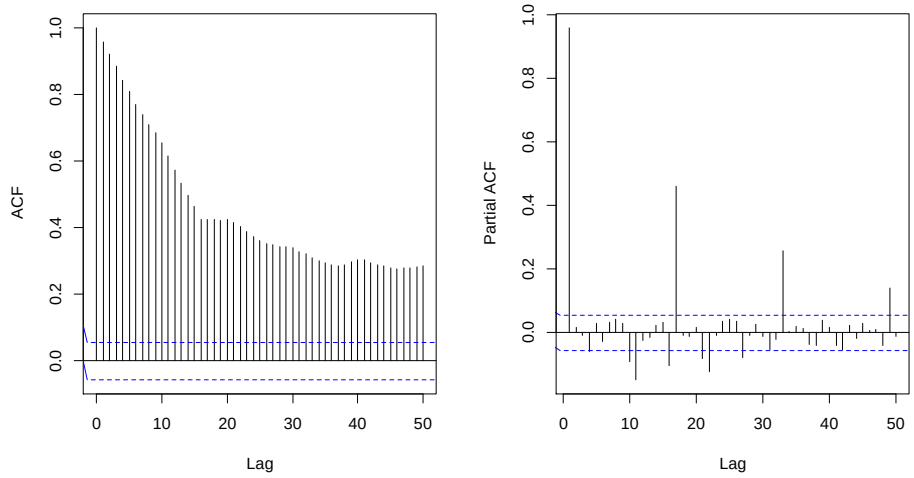


Gráfico 4 - Série Volatilidade NY e série Sentimento.



O Gráfico 5 mostra as autocorrelações e autocorrelações parciais da série Volatilidade NY. Cabe destacar que as observações da série são bastante correlacionadas.

Gráfico 5 - Autocorrelações e autocorrelações parciais da série Volatilidade NY.



A Tabela 43 apresenta os modelos ajustados para os valores das covariáveis diárias, médias (sete e 22 dias) e o modelo mensal da série Volatilidade NY. Em todos os cenários os modelos ajustados foram ARIMA (1, 1, 1). Os coeficientes AR (1) e MA (1) foram significativos em todos os modelos.

No cenário com covariável média 22 dias, houve influência significativa da *Web* diferença (Valor-p=0,028), *Web* o+ (Valor-p=0,008), *Web* o- (Valor-p=0,012), *Web* d+ (Valor-p=0,002) e *Web* d- (Valor-p=0,015) sobre a Volatilidade NY. Já no cenário mensal houve influência significativa da *Web* d+ (Valor-p=0,036) sobre a Volatilidade NY. Nestas situações, à medida que o valor das covariáveis aumenta a Volatilidade NY também aumenta. Nos outros modelos não houve influência significativa das covariáveis sobre a Volatilidade NY.

Tabela 43 - Modelos ARIMA diários, média 7 dias, média 22 dias e mensais para a Volatilidade NY. (Continua)

Coeficientes		M1		M2		M3		M4		M5		M6		M7	
		B	**Valor-p**	**B**	**Valor-p**	**β**	**Valor-p**	**β**	**Valor-p**	**β**	**Valor-p**	**β**	**Valor-p**	**β**	**Valor-p**
Covariável diária	AR (1)	0,96	0,000	0,06	0,904	0,02	0,968	0,05	0,912	0,02	0,966	0,96	0,000	0,03	0,952
	MA (1)	-1,00	0,000	-0,10	0,845	-0,06	0,904	-0,09	0,851	-0,06	0,909	-1,00	0,000	-0,06	0,891
	Sentimento	-	-	0,00	0,890	-	-	-	-	-	-	-	-	-	-
	Web diferença	-	-	-	-	0,00	0,778	-	-	-	-	-	-	-	-
	Web o+ PI	-	-	-	-	-	-	-0,02	0,689	-	-	-	-	-	-
	Web o- PI	-	-	-	-	-	-	-	-	-0,02	0,254	-	-	-	-
	Web d+ PI	-	-	-	-	-	-	-	-	-	-	0,03	0,060	-	-
	Web d- PI	-	-	-	-	-	-	-	-	-	-	-	-	0,02	0,429
Covariável média (7 dias)	AR (1)	0,96	0,000	0,02	0,959	0,01	0,980	0,02	0,967	0,96	0,000	0,04	0,938	0,05	0,909
	MA (1)	-1,00	0,000	-0,06	0,902	-0,05	0,916	-0,06	0,902	-1,00	0,000	-0,07	0,884	-0,09	0,846
	Sentimento	-	-	-0,06	0,248	-	-	-	-	-	-	-	-	-	-
	Web diferença	-	-	-	-	-0,05	0,501	-	-	-	-	-	-	-	-
	Web o+ PI	-	-	-	-	-	-	0,19	0,529	-	-	-	-	-	-
	Web o- PI	-	-	-	-	-	-	-	-	0,13	0,217	-	-	-	-
	Web d+ PI	-	-	-	-	-	-	-	-	-	-	-0,11	0,326	-	-
	Web d- PI	-	-	-	-	-	-	-	-	-	-	-	-	0,28	0,128

Tabela 43 - Modelos ARIMA diários, média 7 dias, média 22 dias e mensais para a Volatilidade NY. (Conclusão)

Coeficientes		M1		M2		M3		M4		M5		M6		M7	
		B	Valor-p	B	Valor-p	β	Valor-p	β	Valor-p	β	Valor-p	β	Valor-p	β	Valor-p
Covariável média (22 dias)	AR (1)	0,96	0,000	0,96	0,000	0,96	0,000	0,05	0,912	0,96	0,000	0,96	0,000	0,04	0,926
	MA (1)	-1,00	0,000	-1,00	0,000	-1,00	0,000	-0,09	0,834	-1,00	0,000	-1,00	0,000	-0,08	0,849
	Sentimento	-	-	0,08	0,618	-	-	-	-	-	-	-	-	-	-
	Web diferença	-	-	-	-	0,51	0,028	-	-	-	-	-	-	-	-
	Web o+ PI	-	-	-	-	-	-	2,61	0,008	-	-	-	-	-	-
	Web o- PI	-	-	-	-	-	-	-	-	0,83	0,012	-	-	-	-
	Web d+ PI	-	-	-	-	-	-	-	-	-	-	1,06	0,002	-	-
	Web d- PI	-	-	-	-	-	-	-	-	-	-	-	-	1,39	0,015
Modelo mensal	AR (1)	0,68	0,000	0,68	0,000	0,70	0,000	0,69	0,000	0,76	0,000	0,65	0,000	0,66	0,000
	MA (1)	-1,00	0,000	-1,00	0,000	-1,00	0,000	-1,00	0,000	-0,99	0,000	-1,00	0,000	-1,00	0,000
	Sentimento	-	-	-0,04	0,937	-	-	-	-	-	-	-	-	-	-
	Web diferença	-	-	-	-	-0,22	0,732	-	-	-	-	-	-	-	-
	Web o+ PI	-	-	-	-	-	-	1,98	0,613	-	-	-	-	-	-
	Web o- PI	-	-	-	-	-	-	-	-	-1,67	0,104	-	-	-	-
	Web d+ PI	-	-	-	-	-	-	-	-	-	-	2,43	0,036	-	-
	Web d- PI	-	-	-	-	-	-	-	-	-	-	-	-	4,20	0,063

Equação do modelo ARIMA (1, 1, 1) sem covariável:

$Y_t = (1 + \varphi_1)Y_{t-1} - \varphi_1 Y_{t-2} - \theta_1(\hat{Y}_{t-1} - Y_{t-1})$,

sendo φ_1 o coeficiente autorregressivo AR (1) e θ_1 o coeficiente média móvel MA (1).

Equação do modelo ARIMA (1, 1, 1) com covariável:

$Y_t = (1 + \varphi_1)Y_{t-1} - \varphi_1 Y_{t-2} - \theta_1(\hat{Y}_{t-1} - Y_{t-1}) + \beta X_{t*}$,

em que β é o coeficiente da covariável e X_{t*} é o valor da covariável no instante *t**: no cenário diário X_{t*} é o próprio valor da covariável referente ao instante *t*, enquanto no cenário médio (sete e 22 dias) X_{t*} é a média dos valores da covariável sete e 22 dias antes do instante t.

Equação do modelo mensal ARIMA (1, 1, 1) sem covariável:

$Y_t = (1 + \varphi_1)Y_{t-1} - \varphi_1 Y_{t-2} - \theta_1(\hat{Y}_{t-1} - Y_{t-1})$,

sendo φ_1 o coeficiente autorregressivo AR (1) e θ_1 o coeficiente médio móvel MA (1).

Equação do modelo mensal ARIMA (1, 1, 1) com covariável:

$Y_t = (1 + \varphi_1)Y_{t-1} - \varphi_1 Y_{t-2} - \theta_1(\hat{Y}_{t-1} - Y_{t-1}) + \beta X_{t-1}$,

em que β é o coeficiente da covariável e X_{t-1} é o valor médio da covariável no mês anterior.

A Tabela 44 mostra a comparação dos modelos sem e com covariáveis em cada cenário. Em nenhum dos cenários, os modelos sem covariáveis apresentaram menores AIC que os modelos com covariáveis. Além disso, observou-se, nos cenários com covariável média de sete e 22 dias, que o modelo com *Web* o- apresentou resultados preditivos melhores que o modelo sem covariável (sete dias: Valor-p = 0,001, 22 dias: Valor-p = 0,028), sendo que o

RMSE, MAD e SMAPE passaram de 1,000; 0,867 e 11,991% para 0,950, 0,825 e 11,406% no modelo com covariável média sete dias, e para 0,971, 0,810 e 11,205% no modelo com covariável média 22 dias, respectivamente.

As previsões atualizadas para 15 dias/5 meses apresentaram melhores resultados preditivos que as previsões simples nos cenários com covariável diária e média (sete e 22 dias), porém no cenário mensal as previsões simples foram melhores que as atualizadas. E, além disso, os modelos mensais foram os que apresentaram as piores medidas de acurácia, indicando que as previsões neste cenário foram piores que nos demais modelos.

Tabela 44 - Comparação dos ajustes dos modelos para a Volatilidade NY. (Continua)

Modelos			AIC	Valor-p[1]	Previsão 15 dias/5 meses			Previsão Atualizada		
					RMSE	MAD	SMAPE	RMSE	MAD	SMAPE
Covariável diária ARIMA (1, 1, 1)	M1	Sem covariáveis	2534,480	-	1,000	0,867	11,991%	0,528	0,259	3,582%
	M2	Sentimento	2555,346	0,932	1,099	0,915	12,653%	0,537	0,258	3,574%
	M3	*Web* diferença	2555,282	0,916	1,096	0,908	12,556%	0,537	0,259	3,581%
	M4	*Web* o+ PI	2555,205	0,932	1,096	0,912	12,618%	0,537	0,259	3,577%
	M5	*Web* o- PI	2554,067	0,940	1,109	0,923	12,762%	0,537	0,258	3,574%
	M6	*Web* d+ PI	2530,904	0,281	0,996	0,841	11,626%	0,523	0,257	3,553%
	M7	*Web* d- PI	2554,736	0,932	1,103	0,918	12,694%	0,537	0,259	3,577%
Covariável média (7 dias) ARIMA (1, 1, 1)	M1	Sem covariáveis	2534,480	-	1,000	0,867	11,991%	0,528	0,259	3,582%
	M2	Sentimento	2554,030	0,932	1,099	0,915	12,650%	0,537	0,258	3,575%
	M3	*Web* diferença	2554,916	0,958	1,104	0,940	13,006%	0,537	0,259	3,579%
	M4	*Web* o+ PI	2554,965	0,932	1,102	0,922	12,756%	0,537	0,259	3,583%
	M5	*Web* o- PI	2532,965	0,001	0,950	0,825	11,406%	0,536	0,268	3,712%
	M6	*Web* d+ PI	2554,397	0,849	1,053	0,921	12,734%	0,537	0,258	3,570%
	M7	*Web* d- PI	2553,038	0,932	1,140	0,944	13,052%	0,537	0,258	3,574%

Tabela 44 - Comparação dos ajustes dos modelos para a Volatilidade NY. (Conclusão)

Modelos			AIC	Valor-p[1]	Previsão 15 dias/5 meses			Previsão Atualizada		
					RMSE	MAD	SMAPE	RMSE	MAD	SMAPE
Covariável média (22 dias) ARIMA (1, 1, 1)	M1	Sem covariáveis	2534,480	-	1,000	0,867	11,991%	0,528	0,259	3,582%
	M2	Sentimento	2534,225	0,076	0,993	0,863	11,939%	0,527	0,259	3,576%
	M3	*Web* diferença	2529,721	0,932	1,350	1,044	14,445%	0,519	0,307	4,251%
	M4	*Web* o+ PI	2548,332	0,991	1,066	0,939	12,989%	0,537	0,260	3,594%
	M5	*Web* o- PI	2528,213	0,028	0,971	0,810	11,205%	0,520	0,316	4,375%
	M6	*Web* d+ PI	2525,090	0,985	1,702	1,362	18,836%	0,518	0,286	3,960%
	M7	*Web* d- PI	2549,413	0,940	1,071	0,911	12,601%	0,537	0,260	3,591%
Modelo mensal ARIMA (1, 1, 1)	M1	Sem covariáveis	206,267	-	0,564	0,390	5,670%	0,594	0,445	6,463%
	M2	Sentimento	201,843	0,594	0,556	0,391	5,683%	0,592	0,443	6,442%
	M3	*Web* diferença	201,733	0,906	0,572	0,403	5,853%	0,593	0,455	6,607%
	M4	*Web* o+ PI	201,594	0,594	0,537	0,398	5,780%	0,586	0,466	6,768%
	M5	*Web* o- PI	199,445	0,969	0,776	0,725	10,538%	0,697	0,635	9,219%
	M6	*Web* d+ PI	197,487	0,219	0,427	0,295	4,280%	0,604	0,452	6,570%
	M7	*Web* d- PI	198,436	0,594	0,705	0,500	7,258%	0,704	0,497	7,219%

[1] Teste de Wilcoxon Pareado.

A análise dos resíduos dos modelos para a Volatilidade NY é apresentada na Tabela 45. Apenas os resíduos dos modelos mensais foram normalmente distribuídos. A suposição de normalidade dos resíduos é essencial em pequenas amostras, pois o Teorema do Limite Central (SHELDON, 2002) não pode ser utilizado para fornecer distribuições aproximadas da Normal. Como neste trabalho, o tamanho amostral é grande e a violação desta suposição não representa um problema.

Em todos os modelos, os resíduos foram independentes (Valores-p Ljung-Box > 0,05).

Tabela 45 - Análise dos resíduos dos modelos para a Volatilidade NY.

(Continua)

Modelo		Jarque-Bera	Ljung-Box (lag 1)	Ljung-Box (lag 2)	Ljung-Box (lag 3)
Covariável diária	M1 Sem covariáveis	0,000	0,579	0,770	0,199
	M2 Sentimento	0,000	1,000	0,994	0,612
	M3 *Web* diferença	0,000	0,998	0,990	0,614
	M4 *Web* o+ PI	0,000	0,999	0,996	0,609
	M5 *Web* o- PI	0,000	0,996	0,984	0,613
	M6 *Web* d+ PI	0,000	0,622	0,788	0,212
	M7 *Web* d- PI	0,000	0,997	0,990	0,601
Covariável média (7 dias)	M1 Sem covariáveis	0,000	0,579	0,770	0,199
	M2 Sentimento	0,000	0,997	0,989	0,630
	M3 *Web* diferença	0,000	0,997	0,987	0,603
	M4 *Web* o+ PI	0,000	0,999	0,984	0,596
	M5 *Web* o- PI	0,000	0,622	0,802	0,215
	M6 *Web* d+ PI	0,000	0,995	0,989	0,621
	M7 *Web* d- PI	0,000	0,999	0,995	0,583

Tabela 45 - Análise dos resíduos dos modelos para a Volatilidade NY.

(Conclusão)

Modelo		Jarque-Bera	Ljung-Box (lag 1)	Ljung-Box (lag 2)	Ljung-Box (lag 3)
Covariável média (22 dias)	M1 Sem covariáveis	0,000	0,579	0,770	0,199
	M2 Sentimento	0,000	0,590	0,769	0,196
	M3 *Web* diferença	0,000	0,614	0,804	0,246
	M4 *Web* o+ PI	0,000	0,991	0,977	0,631
	M5 *Web* o- PI	0,000	0,591	0,782	0,259
	M6 *Web* d+ PI	0,000	0,469	0,719	0,205
	M7 *Web* d- PI	0,000	0,998	0,996	0,617
Modelo mensal	M1 Sem covariáveis	0,768	0,934	0,976	0,996
	M2 Sentimento	0,817	0,931	0,973	0,995
	M3 *Web* diferença	0,803	0,908	0,959	0,992
	M4 *Web* o+ PI	0,828	0,917	0,989	0,999
	M5 *Web* o- PI	0,661	0,673	0,696	0,857
	M6 *Web* d+ PI	0,637	0,941	0,997	0,978
	M7 *Web* d- PI	0,681	0,886	0,976	0,819

No modelo média 22 dias, as variáveis *Web* diferença, *Web* o+, *Web* o-, *Web* d+ e *Web* d- influenciaram significativamente a Volatilidade NY e, no modelo mensal, houve influência significativa da variável *Web* d+. Quanto às previsões simples para 15 dias, nos modelos média sete e 22 dias com *Web* o- obteve-se resultados preditivos superiores aos do modelo sem covariável.

4.4.5 Análise qualitativa dos requisitos de inteligência

Com o objetivo de demonstrar o valor adquirido no processo de análise para IC, pela perspectiva qualitativa, o artefato foi apresentado ao especialista coordenador do BIC que o utilizou para analisar as notícias coletadas no mês de maio, organizadas de acordo com os critérios de classificação para dois

requisitos de inteligência: Aumento da Produção em Rivais e Desenvolvimento da Indústria.

Para comparação, o especialista considerou o resultado apresentado pelo sistema em relação ao Relatório de Tendências do Café publicado no mês de maio pelo BIC. E durante a análise das notícias, respondeu ao questionário com questões descritas na Tabela 18 da Seção 4.4.1.2. O resultado é apresentado por requisito de inteligência nas próximas seções.

4.4.5.1 Aumento da produção em países rivais

As notícias apresentadas ao especialista estão descritas no Anexo A e as respostas do especialista no Anexo B. Pela análise das respostas, pode-se concluir que quanto à Identificação de Necessidades, o protótipo apresentou notícias relevantes e de fontes relevantes frequentemente consultadas e reconhecidas mundialmente. Cabe ressaltar que além destas fontes, notícias relevantes de fontes locais foram capturadas, o que sugere um ganho de escala operacional no que diz respeito a abrangência da pesquisa em relação ao processo manual.

Destacam-se as notícias relevantes que representam ameaças e aumento de produção, referentes a fatores climáticos e iniciativa governamental.

Quanto à Aquisição de Informação Competitiva, para análise SWOT foi relatada imprecisão, pois o sistema, mesmo identificando corretamente uma ameaça, não consegue distinguir que a ameaça não é para o Brasil e sim para outro país, portanto deveria ser classificada como oportunidade ou descartada para análise SWOT. Outro problema é a insuficiência de notícias para Forças e Fraquezas.

Foi possível identificar evidências que impactam oferta e demanda, entretanto o especialista ressalta que a confirmação só pode ser verificada em

médio e longo prazos. Houve incidência de notícias pouco relevantes neste quesito.

O sistema apresentou notícias relevantes sobre a Tanzânia não recuperadas no processo manual, o que sugere capacidade de acrescentar informação competitiva ao processo operacional do BIC e contribuir para a identificação de informações relevantes para a produção de café no mundo.

4.4.5.2 Desenvolvimento da indústria

As notícias apresentadas ao especialista para o requisito Desenvolvimento da Indústria estão descritas no Anexo A e as respostas do especialista no Anexo B. Pela análise das respostas pode-se concluir que quanto à Identificação de Necessidades, o protótipo apresentou notícias relevantes e de fontes relevantes frequentemente consultadas e reconhecidas mundialmente. Também apresentou notícias relevantes para a Indústria, entretanto elas se concentraram em poucos assuntos o que mostra restrição quanto à capacidade de alcance do artefato.

Quanto à Aquisição de Informação Competitiva, para análise SWOT, foi relatada imprecisão e insuficiência de notícias para todas as categorias. Há a sugestão de ampliar a amostra de treinamento para estas categorias, fato que surge pelo envolvimento do especialista no processo de desenvolvimento do sistema.

O sistema contribui para identificar evidências que impactam oferta e demanda de café industrializado e apesar de não acrescentar nova informação ao processo manual no mês de maio de 2016, é útil para agilizar o processo de busca e análise realizado pelos analistas da equipe.

5 CONSIDERAÇÕES FINAIS E LIMITAÇÕES

Na investigação sobre o conceito e aplicação de Mineração Textual para Inteligência Competitiva, foi possível constatar que a literatura é abrangente, mas não inclui explicitamente aplicações na cafeicultura e encontra-se mais contribuições teóricas que práticas. Neste aspecto, a participação de especialistas que lidam diariamente, em organizações do setor, com pesquisa em Inteligência Competitiva para Cafeicultura, forneceu respaldo ao desenvolvimento dos protótipos do TMCIS para pesquisar, extrair da *Web*, classificar e analisar evidências qualitativas em notícias sobre o mercado de café.

O problema delineado e levantamento de requisitos indicam um caráter multidisciplinar em que há critérios de avaliação difundidos na literatura para os aspectos ligados à Computação – avaliação do processo de classificação textual e Economia – análise estatística de séries de preço, entretanto, a lacuna de IC para cafeicultura também impôs como desafio a definição de um modelo de avaliação, uma vez que não era objetivo avaliar o artefato pela perspectiva de qualidade de *software*, mas sim sua contribuição para apoio a IC. Neste ponto a participação dos especialistas foi significativa.

Os artefatos desenvolvidos em dois ciclos da *Design Science Research* foram protótipos para construção do conhecimento para solução do problema proposto nesta pesquisa. Uma vez demonstrados viáveis, serão integrados na construção do sistema para o BIC. O processo mostrou como os especialistas devem estar envolvidos na definição de requisitos e projeto do sistema.

5.1 Limitações

As limitações do trabalho foram divididas de acordo com os módulos de coleta, pré-processamento, classificação e avaliação.

Na coleta, uma limitação é a dependência de um motor de busca específico. Apesar de usar uma ferramenta aprimorada por uma empresa

especializada para busca na *web*, deve-se considerar que a busca fica limitada ao critério de relevância apenas deste fabricante e pode ainda se restringir a preferências de usuário captadas à medida que a ferramenta é utilizada, criando um viés para determinados assuntos preferidos. Outro ponto é a restrição de uso do recurso para o caso de consultas automáticas realizadas por *software*.

A coleta para gerar as séries temporais considerou as 10 primeiras notícias classificadas para cada mês, a partir de 2011, de acordo com o critério de relevância de um motor de busca para cada termo consultado. Portanto, as séries geradas são consideradas simulações em um contexto específico e suscita novas simulações com uma busca em maior profundidade com mais ferramentas de busca. Ao ser integrado, o sistema realizará a coleta diariamente.

Os termos definidos com os especialistas não foram combinados em um contexto mais significativo para a busca na *web*. Por exemplo, as palavras *production* e *Vietnam* foram combinadas apenas com a palavra *coffee*, ao passo que o conjunto *coffee production Vietnam* expressa um contexto mais específico.

O trabalho inclui apenas notícias em Inglês, não incluindo fatos relevantes publicados em outros idiomas.

Todo o conteúdo das páginas recuperadas pela busca é utilizado para gerar o modelo para classificação, o que aumenta o número de palavras com pouco significado nas notícias, não eliminadas como *stopwords*, que compõem o modelo, criando uma matriz muito esparsa para o classificador, o que configura uma limitação na etapa de pré-processamento.

Na classificação de oferta e demanda e fatores de impacto, a amostra de treinamento é pequena, para algumas categorias, desbalanceada e não há uma base de treinamento com notícias irrelevantes. Estas características comprometem a classificação e devem ser tratadas pelos especialistas no módulo de supervisão à medida que o sistema for utilizado.

Para identificação de notícias sobre rivais, foi considerada a ocorrência do nome do país no título das notícias em consulta com os parâmetros de cadeia produtiva, oferta e demanda e fatores de impacto.

A avaliação foi realizada para dois requisitos de inteligência levantados junto aos especialistas, os demais serão avaliadas na continuidade do trabalho. O critério adotado para avaliar qualitativamente foi adaptado da literatura para o contexto pretendido mas é ainda limitado e necessita de mais estudos com os especialistas.

A análise das séries temporais considerou um período específico – 2011 a 2015, o que suscita uma análise de períodos maiores.

6 CONCLUSÃO

Para corroborar que é possível promover Inteligência Competitiva para apoiar decisões sobre gerenciamento de risco e competitividade na cafeicultura por meio de Mineração Textual de notícias publicadas na *web*, foi necessário explorar fenômenos artificiais, portanto, seguindo a metodologia *Design Science Research*, a construção do conhecimento se deu pelo desenvolvimento e avaliação de um sistema baseado em Mineração Textual para Inteligência Competitiva (TMCIS) na cafeicultura.

Um protótipo foi construído em dois ciclos da metodologia, com a participação de especialistas do Bureau de Inteligência do Café, projeto do Centro de Inteligência em Mercado da Universidade Federal de Lavras, gerando os módulos de coleta, pré-processamento, supervisão, classificação e análise para desenvolvimento do TMCIS.Este processo ampliou o conhecimento sobre IC para cafeicultura e potencial da Mineração Textual para lidar com dados da *web*. As principais conclusões são relatadas nesta seção na sequência dos objetivos propostos.

Ao investigar o conceito e a aplicação de TMCIS, para apoio à obtenção de Inteligência Competitiva, concluiu-se que é necessária a definição de um processo de IC, que por sua vez exige o levantamento de requisitos de inteligência para o domínio específico e quais informações devem ser extraídas dos textos. Assim adotou-se um modelo geral de processo de IC e por meio de pesquisa com especialistas, foram propostos oito requisitos de inteligência para a cafeicultura e evidências qualitativas em notícias sobre o mercado de café que as classificam em categorias relevantes para identificar os requisitos de inteligência propostos.

Durante o desenvolvimento dos módulos para extrair da *web*, classificar e analisar as evidências qualitativas, concluiu-se que a etapa de consulta e pré-processamento são determinantes para a qualidade da classificação. No caso da

coleta, o conjunto de palavras significativas para a cafeicultura e a combinação delas é um fator importante para que a busca, via motores de busca, retorne notícias pertinentes ao contexto. Foi proposto um conjunto de palavras definido juntamente com os especialistas considerando os requisitos de inteligência, cadeia produtiva, forças competitivas e fatores que impactam oferta e demanda de café.

Na etapa de pré-processamento, o desafio foi a definição de contexto para a abordagem de classificação textual por aprendizado de máquina. A criação de uma base de treinamento exclusivamente por classificação manual, além de custosa, está sujeita a inserção de ruídos que comprometem o processo, portanto o módulo de Supervisão deve ser aprimorado para auxiliar o desenvolvimento de um dicionário de termos (Ontologia) ou dicionário léxico para cafeicultura e combinar as abordagens de orientação semântica e classificação por aprendizado de máquina.

A classificação quanto aos setores da cadeia produtiva pré-definidos pelos especialistas, realizada no módulo de coleta, é útil para a seleção de notícias relevantes e organização da informação para análises nas categorias Indústria e Produção, para as demais categorias, é necessário melhorar a base de treinamento com amostras mais representativas.

A amostra utilizada para treinamento dos classificadores compromete as séries de ocorrências de notícias e deve ser melhorada para obter melhor desempenho com os classificadores. É necessária a criação de uma amostra de treinamento mais significativa. Espera-se que com a implantação do sistema e utilização do Módulo de Supervisão esta tarefa seja realizada.

Ao verificar em que medida a informação qualitativa coletada da *web* apresenta correlação com a variação de preço do café na Bolsa de Nova York, pode-se confirmar com abordagem estatística, a eficiência do mercado de café, como esperado – não se pode observar relação causal entre a ocorrência de

notícias e preço, entretanto, mesmo com as limitações apontadas para a classificação, observa-se uma correlação no caso da média de 22 períodos de notícias sobre oferta e demanda e volatilidade do preço na bolsa de NY, que sugere mais estudos em busca de um indicador que configure um cenário de possível aumento de volatilidade.

Outra conclusão é a evidência de correlação entre notícias da produção e notícias sobre oferta e preço mensal, o que corrobora estudos como os de Martins (2015) que apontam influência da produção e fatores climáticos no mercado futuro.

Apesar das limitações, o resultado mostra a viabilidade de mais estudos em busca de um indicador mais preciso que promova o delineamento de um cenário de oferta e demanda para decisões na cafeicultura. Entretanto, a eficiência do mercado e a complexidade para eliminar ruídos e incidência de erros, indica que tal indicador quantitativo de evidências qualitativas em notícias deve ser avaliado por especialistas juntamente com dados da Análise Fundamentalista para tomada de decisões.

Ao verificar a contribuição do artefato para apoiar Inteligência Competitiva no BIC, a avaliação qualitativa do especialista mostra que o sistema, mesmo com limitações, auxilia a análise do requisito de inteligência Aumento de Produção em Países Rivais e Desenvolvimento da Indústria, mas exige melhorias para aumentar precisão e abrangência.

As avaliações mostram que o sistema permite desenvolver capacidades dinâmicas contribuindo em nível de Estratégia Funcional, para organizações do setor cafeeiro, uma vez que o resultado sugere um cenário a partir de dados coletados da *web*, no qual o tomador de decisão pode optar por realizar ou não o *hedge* para gerenciamento de risco.

Entretanto são necessários mais estudos, desenvolvimento e avaliações para melhorar a precisão em busca de mais segurança sobre o momento de

realizar *hedge*. Neste caso, com o conhecimento adquirido após duas etapas da metodologia, não se pode esperar um indicador quantitativo preciso para operações em curto prazo, mas sim um alerta de evidências qualitativas vindas da *web* que, a saber, sua influência e erro, e confrontado com uma análise qualitativa, permita perceber aumento de volatilidade e viés para delinear um cenário de tomada de decisões.

Pode-se concluir, pelos métodos de avaliação utilizados e conhecimento adquirido até o segundo ciclo da metodologia, que é possível obter Inteligência Competitiva para gerenciamento de risco e decisões na cafeicultura a partir de notícias publicadas na *web* através de um sistema baseado em Mineração Textual, pois, apesar das limitações, a avaliação apresenta as seguintes evidências:

a) Houve correlação entre notícias classificadas como categoria produção e preço do café NY no período estudado;
b) A série gerada a partir das notícias de oferta e demanda contribuiu, mesmo que de forma pouco significativa, para explicação do preço no período anterior a fevereiro de 2014 e posterior, pela abordagem adotada;
c) É plausível obter um indicador a partir de uma série cronológica de eventos com evidências qualitativas sobre oferta e demanda em notícias relacionado a volatilidade de preço do café;
d) O artefato proposto acrescenta a capacidade de recuperar e organizar um conjunto relevante de notícias a partir da *web* (volume) para monitorar a produção de café em países rivais, contribuindo para a análise de um requisito de Inteligência definido por especialistas.

A experiência adquirida no desenvolvimento desta pesquisa mostra viabilidade de estudar, pela perspectiva de IC, fatores que impactam a volatilidade e níveis de preço no mercado de café e a possibilidade de incluí-los em modelos de previsão para gerenciamento de risco, mas ao mesmo tempo, explicita as dificuldades e desafios da abordagem com Mineração Textual em grandes volumes de dados e reforça que não há pretensão em esgotar o tema, mas sim necessidade de mais estudos.

6.1 Trabalhos futuros

A continuação do trabalho implica em transformar os protótipos em um sistema integrado aos processos do BIC. Paralelamente, para tratar as limitações citadas neste trabalho e aprimorar o desempenho, as seguintes tarefas são necessárias:

a) Acrescentar novos recursos para a coleta de notícias: mais motores de busca e *web crawlers* a partir de sites específicos;
b) Incluir e comparar a relevância de notícias selecionadas em *sites* de agências especializadas no mercado de café e selecionadas pelos especialistas e classificadas quanto ao número de notícias relevantes validadas existentes no banco de dados;
c) Realizar a coleta de notícias de forma distribuída e diariamente, para aumentar a abrangência, profundidade e variedade sem prejudicar o desempenho do sistema e evitando viés de motores de busca de específicos;
d) Desenvolver uma base de notícias irrelevantes para aprimorar o desempenho do classificador e diminuir a incidência de notícias que não contribuem para análise na base de dados;

e) Aprimorar a etapa de limpeza de texto no pré-processamento para minimizar o ruído do conteúdo coletado;
f) Utilizar o módulo de supervisão para aumentar a base de treinamento de notícias para as categorias propostas: Fatores de Impacto, Oferta e Demanda e SWOT;
g) Comparar uma abordagem por orientação semântica (dicionário léxico) com a classificação supervisionada por aprendizado de máquina, para tal desenvolver um dicionário léxico baseado em uma Ontologia da Cafeicultura para melhorar o conjunto de *stopwords* e organizar a informação de forma mais eficiente para criação de contexto, coleta, classificação e análise;
h) Incluir a possibilidade de usar outros classificadores;
i) Explorar o uso de NLP – *Natural Language Processing* para identificar as entidades que representam as forças de Porter de forma mais precisa, não só no título, mas no corpo do texto, além de relacionamentos entre elas;
j) Estender o sistema para notícias em outras línguas;
k) Acrescentar a análise de outras fontes de informação além das notícias, como relatórios de empresas e opiniões de analistas;
l) Após integrar a primeira versão do sistema ao BIC, realizar a avaliação do sistema segundo o modelo de avaliação de *software* para Inteligência Competitiva e melhorar a usabilidade da interface com o usuário

REFERÊNCIAS

ABREU, G. F. de et al. Identificação das principais tendências para a produção mundial de café. In: SIMPÓSIO DE PESQUISA DOS CAFÉS DO BRASIL, 8., 2013, Salvador. **Anais**... [S.l.]: Consórcio Pesquisa Café, 2013.

AGUWA, C. C.; MONPLAISIR, L.; TURGUT, O. Voice of the customer: customer satisfaction ratio based analysis. **Expert Systems with Applications**, New York, v. 39, n. 11, p. 10112-10119, Sept. 2012.

AKEN, J. E. V. Management research based on the paradigm of the design sciences: the quest for field-tested and grounded technological rules. **Journal of management studies**, Oxford, v. 41, n. 2, p. 219-246, Mar. 2004.

ALLAN, J. (Ed.). **Topic detection and tracking**: event-based information organization. New York: Springer, 2012.

ALMEIDA, R. P. D. O **Comportamento manada em mercados acionários latino-americanos**. 2011. 76 p. Dissertação (Mestrado em Administração)-Universidade Federal de Santa Catarina, Florianópolis, 2011.

ANICA-POPA, I.; CUCUI, G. A framework for enhancing competitive intelligence capabilities using decision support system based on web mining techniques. **International Journal of Computers, Communications and Control**, Oradea, v. 4, n. 4, p. 326-334, 2009.

ANTWEILER, W.; FRANK, M. Z. Is all that talk just noise? The information content of internet stock message boards. **The Journal of Finance**, New York, v. 59, n. 3, p. 1259-1294, June 2004.

ARONSON, D. **Evidence-based technical analysis**: applying the scientific method and statistical inference to trading signals. New York: J. Wiley, 2011.

BAARS, H.; KEMPER, H.-G. Management support with structured and unstructured data—an integrated business intelligence framework. **Information Systems Management**, Boston, v. 25, n. 2, p. 132-148, Apr. 2008.

BACCIANELLA, S.; ESULI, A.; SEBASTIANI, F. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: INTERNATIONAL LANGUAGE RESOURCES AND EVALUATION, 7., 2010, Malta. **Proceedings**… [Luxembourg]: ELRA, 2010. p. 2200-2204.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Recuperação de Informação**: conceitos e tecnologia das máquinas de busca. Porto Alegre: Bookman, 2013.

BAKHT, M. N.; EL-DIRABY, T. E. Synthesis of decision-making research in construction. **Journal of Construction Engineering and Management**, New York, v. 141, n. 9, Sept. 2015.

BALAHUR, A. et al. Sentiment analysis in the news. In: INTERNATIONAL LANGUAGE RESOURCES AND EVALUATION, 7., 2010, Malta. **Proceedings**… [Luxembourg]: ELRA, 2010. p. 2216-2220.

BARDIN, L. **Análise de conteúdo**. [3. ed.]. Lisboa: Edições 70, 2006.

BARROS, Á. D. M.; AGUIAR, D. R. Gestão do risco de preço de café arábica: uma análise por meio do comportamento da base. **Revista de Economia e Sociologia Rural**, Brasília, v. 43, n. 3, p. 443-464, jul./set. 2005.

BAYAZIT, N. Investigating design: a review of forty years of design research. **Design issues**, Chicago, v. 20, n. 1, p. 16-29, Dec. 2004.

BECK, K. **Extreme programming explained**: embrace change. Boston: Addison-Wesley Longman, 2000.

BM&FBOVESPA. **Contrato Futuro de Café Arábica 4/5**. 2015a. Disponível em: <http://www.bmfbovespa.com.br/lumis/portal/file/fileDownload.jsp?fileId=8AA8D097528574830152A1A90CC70CBD>. Acesso em: 20 abr. 2016.

_____. **Contratos Derivativos Futuro de Café Arábica 6/7**. 2015b. Disponível em: <http://www.bmfbovespa.com.br/lumis/portal/file/fileDownload.jsp?fileId=8A828D29514A326701514A54124954E7>. Acesso em: 20 abr. 2016.

BOLLEN, J.; MAO, H.; PEPE, A. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In: INTERNATIONAL CONFERENCE ON WEBLOGS AND SOCIAL MEDIA, 5., 2011, Barcelona. **Proceedings**… Palo Alto: AAAI, 2011. p. 450-453.

BOLLERSLEV, T. Glossary to arch (garch). **CREATES Research Paper**, Aarhus, v. 49, Sept. 2008.

BOSE, R. Competitive intelligence process and tools for intelligence analysis. **Industrial Management and Data Systems**, Wembley, v. 108, n. 4, p. 510-528, Apr. 2008.

_____. Advanced analytics: opportunities and challenges. **Industrial Management and Data Systems**, Wembley, v. 109, n. 2, p. 155-172, Feb. 2009.

BOUTHILLIER, F.; SHEARER, K. **Assessing competitive intelligence software**: a guide to evaluating CI technology. New Jersey: Information Today, 2003.

BOX, G. E. et al. **Time series analysis: forecasting and control**. New York: J. Wiley, 2015.

BRAMER, M. **Principles of data mining**. New York: Springer, 2013.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. **Culturas-café**: saiba mais. [2016?]. Disponível em: <http://www.agricultura.gov.br/vegetal/culturas/cafe/saiba-mais>. Acesso em: 10 mar. 2016.

BRITT, P. The new competitive intelligence: raising the confidence quotient. **KMWorld**, [S.l.], v. 15, n. 10, Nov./Dec, 2006. Disponível em: <http://www.kmworld.com/Articles/Editorial/Features/The-new-competitive-intelligence-Raising-the-confidence-quotient-18527.aspx>. Acesso em: 20 abr. 2016.

BROCK, W.; LAKONISHOK, J.; LEBARON, B. Simple technical trading rules and the stochastic properties of stock returns. **The Journal of Finance**, New York, v. 47, n. 5, p. 1731-1764, Dec. 1992.

BRODY, R. Issues in defining competitive intelligence: an exploration. **IEEE Engineering Management Review**, New York, v. 3, n. 36, p. 3-3, Jan. 2008.

CALOF, J.; RICHARDS, G.; SMITH, J. Foresight, competitive intelligence and business analytics: tools for making industrial programmes more efficient. **Foresight-Russia**, Moscow, v. 9, n. 1, p. 68-81, Mar. 2015.

CARVALHO, A. M. et al. Correlação entre crescimento e produtividade de cultivares de café em diferentes regiões de Minas Gerais, Brasil. **Pesquisa Agropecuária Brasileira**, Brasília, v. 45, n. 3, p. 269-275, mar. 2010. ~~2011~~.

CARVALHO, G. R. **Avaliação de sistemas de pro dução de café na região sul de Minas Gerais**: um modelo de análise de decisão. 2002. 68 p. Dissertação (Mestrado em Ciências)-Universidade de São Paulo, Piracicaba, 2002.

CASTRO JÚNIOR, L. G. **Análise de Mercado, Mercado de Opções e CPR**. Lavras: UFLA, 2008.

CHEN, H.; CHAU, M.; ZENG, D. CI Spider: a tool for competitive intelligence on the Web. **Decision Support Systems**, Amsterdam, v. 34, n. 1, p. 1-17, Dec. 2002.

CHEN, H.; CHIANG, R. H.; STOREY, V. C. Business Intelligence and Analytics: From Big Data to Big Impact. **MIS quarterly**, Minneapolis, v. 36, n. 4, p. 1165-1188, Dec. 2012.

CHEN, K.-Y.; LUESUKPRASERT, L.; CHOU, S.-C. T. Hot topic extraction based on timeline analysis and multidimensional sentence modeling. **IEEE Transactions on Knowledge and Data Engineering**, New York, v. 19, n. 8, p. 1016-1025, Aug. 2007.

CHOI, Y. et al. Identifying sources of opinions with conditional random fields and extraction patterns. In: CONFERENCE ON HUMAN LANGUAGE TECHNOLOGY AND EMPIRICAL METHODS IN NATURAL LANGUAGE, 1., 2005, Stroudsburg. **Proceedings**… Stroudsburg: Association for Computational Linguistics, 2005. p. 355-362.

CHOUDER, M. L.; CHALAL, R. Models and tools support to the Competitive Intelligence process. In: INTERNATIONAL SYMPOSIUM OF KNOWLEDGE MANAGEMENTE, 4., 2014, Maghreb. **Proceedings**… New York: IEEE, 2014.

CHUNG, W. BizPro: Extracting and categorizing business intelligence factors from textual news articles. **International Journal of Information Management**, Guildford, v. 34, n. 2, p. 272-284, 2014.

COLLIER, N. et al. BioCaster: Detecting public health rumors with a Web-based text mining system. **Bioinformatics**, [S.l.], v. 24, n. 24, p. 2940-2941, Dec. 2008. Disponível em: <http://www.scopus.com/inward/record.url?eid=2-s2.0-57249114504&partnerID=40&md5=4fb9462c189c34616bfafb9e4347d11d>. Acesso em: 20 abr. 2016.

CONT, R.; BOUCHAUD, J.-P. Herd behavior and aggregate fluctuations in financial markets. **Macroeconomic dynamics**, New York, v. 4, n. 2, p. 170-196, 2000.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine learning**, Boston, v. 20, n. 3, p. 273-297, 1995.

CRUZ, D. F. et al. Inteligência competitiva em organizações de serviços: uma revisão sistemática da literatura. **Revista Produção Online**, Florianópolis, v. 15, n. 1, p. 50-77, Mar. 2015.

CRUZ, F. L. et al. Building layered, multilingual sentiment lexicons at synset and lemma levels. **Expert Systems with Applications**, New York, v. 41, n. 13, p. 5984-5994, Oct. 2014.

DAI, Y. **Designing text mining-based competitive intelligence systems**. 2013. 136 p. Dissertation (Master in Forestry and Natural Sciences)-University of Eastern Finland, Joensuu, 2013.

DAI, Y. et al. MOETA: a novel text-mining model for collecting and analysing competitive intelligence. **International Journal of Advanced Media and Communication**, [Geneva], v. 5, n. 1, p. 19-39, May 2013.

DAI, Y.; KAKKONEN, T.; SUTINEN, E. MinerVA: a decision support model that uses novel text mining technologies. In: INTERNATIONAL CONFERENCE ON MANAGEMENT AND SERVICE SCIENCE, 4., 2010, Wuhan. **Proceedings**… New York: IEEE, 2010. p. 1-4.

_____. 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**, [Geneva], v. 3, n. 1, p. 165-173, 2011a.

_____. SoMEST: a model for detecting competitive intelligence from social media. In: INTERNATIONAL ACADEMIC MINDTREK CONFERENCE, 15., 2011, Tampere. **Proceedings**… New York: ACM, 2011b.

DAMODARAN, A. **Mitos de Investimentos**. New Jersey: Prentice-Hall, 2006.

DAS, S. R.; CHEN, M. Y. Yahoo! for Amazon: Sentiment extraction from small talk on the web. **Management Science**, Providence, v. 53, n. 9, p. 1375-1388, Sept. 2007.

DAVE, K.; LAWRENCE, S.; PENNOCK, D. M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 12., 2003, Budapest. **Proceedings**… New York: ACM, 2003. p. 519-528.

DAYAL, U. et al. Data integration flows for business intelligence. In: INTERNATIONAL CONFERENCE ON EXTENDING DATABASE TECHNOLOGY, 12., 2009, Saint Petersburg. **Proceedings**… New York: ACM, 2009. p. 1-11.

DEMPSTER, A. P. et al. Maximum likelihood from incomplete data via the EM algorithm. **Journal of the Royal Statistical Society-Series B (methodological)**, [S.l.], p. 1-38, 1977.

DEY, L.; HAQUE, S. M. Opinion mining from noisy text data. **International Journal on Document Analysis and Recognition**, [New York], v. 12, n. 3, p. 205-226, Sept. 2009.

DING, X.; LIU, B. The utility of linguistic rules in opinion mining. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE, 30., 2007, Amsterdam. **Proceedings**… New York: ACM, 2007.

DOU, W. et al. Leadline: Interactive visual analysis of text data through event identification and exploration. In: CONFERENCE ON VISUAL ANALYTICS SCIENCE AND TECHNOLOGY, 17., 2012, Seattle. **Proceedings**… Danvers: IEEE, 2012.

DRESCH, A. **Design science e design science research como artefatos metodológicos para engenharia de produção**. 2013. 184 f. Dissertação (Mestrado em Engenharia de Produção e Sistemas)-Universidade do Vale do Rio dos Sinos, São Leopoldo, 2014.

DU TOIT, A. Comparative Study of Competitive Intelligence Practices between Two Retail Banks in Brazil and South Africa. **Journal of Intelligence Studies in Business**, Halmstad, v. 3, n. 2, p. 30-39, 2013.

DU TOIT, A.; MULLER, M.-L. Organizational structure of competitive intelligence activities: a South African case study. **South African Journal of Information Management**, Durbanville, v. 6, n. 3, Sept. 2004.

DUBE, O.; VARGAS, J. F. Commodity price shocks and civil conflict: Evidence from Colombia. **The Review of Economic Studies**, Oxford, v. 80, n. 4, p. 1384-1421, 2013.

ENGLE, R. F. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. **Econometrica**, Chicago, v. 50, n. 4, p. 987-1007, July 1982.

FAMA, E. F. Efficient capital markets: a review of theory and empirical work. **The journal of Finance**, New York, v. 25, n. 2, p. 383-417, May 1970.

FARINA, E. M. M. Q.; ZYLBERSZTAJN, D. **Competitividade no agribusiness brasileiro**. São Paulo: PENSA/FIA/FEA/USP, 1998. v. 4.

FAYYAD, U. M. et al. **Advances in knowledge discovery and data mining**. Menlo Park: American Association for Artificial Intelligence, 1996.

FELLBAUM, C. **WordNet**. New York: Wiley Online Library, 1998.

FEUERRIEGEL, S.; NEUMANN, D. News or Noise? How News Drives Commodity Prices. In: INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS, 34., 2013, Milan, **Proceedings**… [S.l.]: AIS, 2013.

FILENI, D. H. **O risco de base, a efetividade do hedging e um modelo para a estimativa da base**: uma contribuição ao agronegócio do café em Minas Gerais. 1999. 137 p. Dissertação (Mestrado em Administração Rural)-Universidade Federal de Lavras, Lavras, 1999.

FLEISHER, C. S.; BENSOUSSAN, B. E. **Business and competitive analysis**: effective application of new and classic methods. New York: Pearson Education, 2014.

FONTES, R. E.; CASTRO JUNIOR, L. G.; AZEVEDO, A. F. Base e risco de base da cafeicultura em Minas Gerais e São Paulo. **Resenha BM&F**, São Paulo, n. 153, p. 50-56, 2003.

FRY, J. M.; LAI, B.; RHODES, M. The interdependence of coffee spot and futures markets. **International Network for Economic Research Working Paper Series**, London, v. 1, 2011.

GARCÍA-CUMBRERAS, M. Á.; MONTEJO-RÁEZ, A.; DÍAZ-GALIANO, M. C. Pessimists and optimists: improving collaborative filtering through sentiment analysis. **Expert Systems With Applications**, New York, v. 40, n. 17, p. 6758-6765, Dec. 2013.

GEROW, A.; KEANE, M. T. Mining the web for the" voice of the herd" to track stock market bubbles. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 22., 2011, Barcelona. **Proceedings**… Palo Alto: AAAI, 2011.

GOMES, E.; BRAGA, F. **Inteligência competitiva**: como transformar informação em um negócio lucrativo. São Paulo: Elsevier, 2004.

GONÇALVES, P. et al. Comparing and combining sentiment analysis methods. In: CONFERENCE ON ONLINE SOCIAL NETWORK, 1., 2013, Boston. **Proceedings**… New York: ACM, 2013. p. 27-38.

GÖRENER, A.; TOKER, K.; ULUÇAY, K. Application of combined SWOT and AHP: a case study for a manufacturing firm. **Procedia-Social and Behavioral Sciences**, [Amsterdam], v. 58, p. 1525-1534, Oct. 2012. ISSN 1877-0428.

GRIMMER, J.; STEWART, B. M. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. **Political Analysis**, Oxford, v. 21, n. 3, p. 267-297, Jan. 2013..

GROSSI, J. Administrar o agronegócio do café é o maior desafio. **Preços Agrícolas**, Piracicaba, v. 12, n. 142, p. 8-8, ago. 1998.

GUIMARÃES, F. dos R. **Descobrindo padrões de gêneros das mensagens em fóruns de discussão de ambientes virtuais de aprendizagem via mineração de texto**. 2015. 114 p. Dissertação (Mestrado em Ciência da Computação)- Universidade Federal de Lavras, Lavras, 2015.

GUPTA, V.; LEHAL, G. S. A survey of text mining techniques and applications. **Journal of emerging technologies in web intelligence**, [S.l.], v. 1, n. 1, p. 60-76, Aug. 2009.

HADDI, E.; LIU, X.; SHI, Y. The role of text pre-processing in sentiment analysis. **Procedia Computer Science**, [Amsterdam], v. 17, p. 26-32, 2013.

HEARST, M. A. et al. Support vector machines. **IEEE Intelligent Systems and their Applications**, New York, v. 13, n. 4, p. 18-28, 1998.

HERRING, J. P. Key intelligence topics: a process to identify and define intelligence needs. **Competitive Intelligence Review**, New York, v. 10, n. 2, p. 4-14, 2nd Quarter 1999.

HEUMESSER, C.; STARITZ, C. Financialisation and the microstructure of commodity markets: a qualitative investigation of trading strategies of financial investors and commercial traders. **ÖFSE**, Vienna, Oct. 2013. Working Paper 44.

HILL, T.; WESTBROOK, R. SWOT analysis: it's time for a product recall. **Long Range Planning**, London, v. 30, n. 1, p. 46-52, Feb. 1997.

HOHHOF, B. Developing information systems for competitive intelligence support. **Library Trends**, Baltimore, v. 43, n. 2, p. 226-238, Fall 1994.

HOLLANDER, M.; WOLFE, D. A.; CHICKEN, E. **Nonparametric statistical methods**. New York: J. Wiley, 2013.

HOLMES, G.; DONKIN, A.; WITTEN, I. H. Weka: a machine learning workbench. In: AUSTRALIAN AND NEW ZEALAND CONFERENCE ON INTELLIGENT INFORMATION SYSTEMS, 2., 1994, Brisbane. **Proceedings**… [S.l.: s.n.], 1994. p. 357-361.

HU, M.; LIU, B. Mining and summarizing customer reviews. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 10., 2004, Seattle. **Proceedings**… New York: ACM, 2004. p. 168-177.

HUANG, D. et al. Discovering event evolution graphs based on news articles relationships. In: INTERNATIONAL CONFERENCE ON E-BUSINESS ENGINEERING, 11., 2014, Guangzhou. **Proceedings**… New York: IEEE, 2014. p. 246-251.

HUANG, K.-W.; LI, Z. A multilabel text classification algorithm for labeling risk factors in SEC form 10-K. **ACM Transactions on Management Information Systems**, New York, v. 2, n. 3, p. 18-39, Aug. 2011.

HULL, J. **Fundamentos dos mercados futuros e de opções**. São Paulo: Bolsa de Mercadorias & Futuros, 2005.

HUMPHERYS, S. L. et al. Identification of fraudulent financial statements using linguistic credibility analysis. **Decision Support Systems**, Amsterdam, v. 50, n. 3, p. 585-594, Feb. 2011.

JARQUE, C. M.; BERA, A. K. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. **Economics letters**, Amsterdam, v. 6, n. 3, p. 255-259, 1980.

JOACHIMS, T. **Text categorization with support vector machines: Learning with many relevant features**. New York: Springer, 1998.

JOHN, G. H.; LANGLEY, P. Estimating continuous distributions in Bayesian classifiers. In: CONFERENCE ON UNCERTAINTY IN ARTIFICIAL INTELLIGENCE, 11., 1995, Montreal. **Proceedings**… [Burlington: Morgan Kaufmann Publishers], 1995. p. 338-345.

JOHNSON, G.; SCHOLES, K.; WHITTINGTON, R. 8. ed. **Exploring corporate strategy**: text and cases. New Jersey: Pearson Hall, 2008.

JORDAN, A. On discriminative vs. generative classifiers: a comparison of logistic regression and naive bayes. **Advances in Neural Information Processing Systems**, San Mateo, v. 14, p. 841-849, 2002.

JUNQUÉ DE FORTUNY, E. et al. Media coverage in times of political crisis: A text mining approach. **Expert Systems with Applications**, New York, v. 39, n. 14, p. 11616-11622, Oct. 2012. Disponível em: <http://www.scopus.com/inward/record.url?eid=2-s2.0-84861836775&partnerID=40&md5=55ea54eca6bffa8c534f3c7d6e768986>. Acesso em: 20 abr. 2016.

KAHANER, L. **Competitive intelligence**: how to gather analyze and use information to move your business to the top. New York: Simon and Schuster, 1997.

KAHNEMAN, D.; TVERSKY, A. Prospect theory: An analysis of decision under risk. **Econometrica**, Chicago, v. 47, n. 2, p. 263-291, Mar. 1979.

KAMPS, J. et al. Using wordnet to measure semantic orientations of adjectives. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 4. 2004, Lisbon. **Proceedings**… [Luxembourg]: ELRA, 2004.

KANG, H.; YOO, S. J.; HAN, D. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. **Expert Systems with Applications**, New York, v. 39, n. 5, p. 6000-6010, Apr. 2012. Disponível em: <http://www.scopus.com/inward/record.url?eid=2-s2.0-84855887855&partnerID=40&md5=b01a78e61d8a6b6a4f544f9017e26597>. Acesso em: 20 abr. 2016.

KANTARDZIC, M. **Data mining**: concepts, models, methods, and algorithms. New York: J. Wiley, 2011.

KRAVET, T.; MUSLU, V. Textual risk disclosures and investors' risk perceptions. **Review of Accounting Studies**, Philadelphia, v. 18, n. 4, p. 1088-1122, Dec. 2013.

KUCKARTZ, U. QDA-Software im Methodendiskurs: Geschichte, Potenziale, Effekte. In: KUCKARTZ, Y.; GRUNENBERG, H.; DRESING, T. (Ed.). **Qualitative Datenanalyse**: computergestützt. New York: Springer, 2004. p.11-26.

KURBY, C. A.; ZACKS, J. M. Segmentation in the perception and memory of events. **Trends in Cognitive Sciences**, Cambridge, Estados Unidos, v. 12, n. 2, p. 72-79, Feb. 2008.

KUTCHUKIAN, E. **O efeito manada nos fundos de investimento no Brasil**: um teste em finanças comportamentais. 2010. 58 f. Dissertação (Mestrado em Administração)-Fundação Getúlio, São Paulo, 2010.

LACERDA, D. P. et al. Design Science Research: método de pesquisa para a engenharia de produção. **Gestão & Produção**, São Carlos, v. 20, n. 4, p. 741-761, 2013.

LE MOIGNE, J.-L. O construtivismo: dos fundamentos. Lisboa: Instituto Piaget, 1994. v. 1.

LEE, C. J.; WU, Y. C.; CHEN, Y. C. Building news sentiment indicators for stock marketing application. **International Journal of Advancements in Computing Technology**, [Seoul], v. 4, n. 2, p. 103-110, Feb. 2012. Disponível em: <http://www.scopus.com/inward/record.url?eid=2-s2.0-84863183789&partnerID=40&md5=06c1b4831a52fad67340c61e70268262>. Acesso em: 20 abr. 2016.

LEWIS, D. D. et al. Rcv1: A new benchmark collection for text categorization research. **The Journal of Machine Learning Research**, [S.l.], v. 5, p. 361-397, Dec. 2004.

LI, F. The information content of forward-looking statements in corporate filings: a naïve Bayesian machine learning approach. **Journal of Accounting Research**, Chicago, v. 48, n. 5, p. 1049-1102, Dec. 2010.

LI, N. et al. Network environment and financial risk using machine learning and sentiment analysis. **Human and Ecological Risk Assessment**, Amherst, v. 15, n. 2, p. 227-252, Mar. 2009. Disponível em: <http://www.scopus.com/inward/record.url?eid=2-s2.0-68749112629&partnerID=40&md5=21716f726c8e623dc861a480052ad59d>. Acesso em: 20 abr. 2016.

LIMA JÚNIOR, P. de O.; CASTRO JÚNIOR, L. G. de; ZAMBALDE, A. L. Avaliação de métodos de classificação textual para apoio a análise de conteúdo aplicada a gestão da informação no mercado de café. In: SEMINÁRIOS EM ADMINISTRAÇÃo, 18., 2015, São Paulo. Anais… São Paulo: USP, 2015. Disponível em: <http://sistema.semead.com.br/18semead/resultado/trabalhosPDF/929.pdf>. Acesso em: 20 abr. 2016.

LIM, E.-P.; CHEN, H.; CHEN, G. Business intelligence and analytics: research directions. **ACM Transactions on Management Information Systems**, New York, v. 3, n. 4, p. 1-10, Jan. 2013.

LIMA JÚNIOR, P.; CASTRO JÚNIOR, L.; ZAMBALDE, A. Analysis of machine learning techniques to classify news for information management in coffee market. **IEEE Latin America Transactions**, New York, v. 13, n. 7, p. 2285-2291, July 2015.

LIU, B. Sentiment analysis: a multi-faceted problem. **IEEE Intelligent Systems**, New York, v. 25, n. 3, p. 76-80, 2010.

LIU, B.; HU, M.; CHENG, J. Opinion observer: analyzing and comparing opinions on the web. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 14., 2005, Chiba. **Proceedings**… New York: ACM, 2005. p. 342-351.

LIU, Y. et al. Identifying helpful online reviews: a product designer's perspective. **Computer-Aided Design**, Guildford, v. 45, n. 2, p. 180-194, Feb., 2013.

LJUNG, G. M.; BOX, G. E. On a measure of lack of fit in time series models. **Biometrika**, London, v. 65, n. 2, p. 297-303, 1978.

LLOYD, L.; KECHAGIAS, D.; SKIENA, S. Lydia: a system for large-scale news analysis. In: CONSENS, M.; NAVARRO, G. (Ed.). **String processing and information retrieval**. Berlin: Springer, 2005. p. 161-166.

LO, A. W.; MAMAYSKY, H.; WANG, J. Foundations of technical analysis: computational algorithms, statistical inference, and empirical implementation. **The Journal of Finance**, New York, v. 55, n. 4, p. 1705-1770, Aug. 2000.

LOPES, T. J. P. et al. Mineração de opiniões aplicada à análise de investimentos. In: BRAZILIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB, 14., 2008, Vila Velha. **Proceedings**… New York: ACM, 2008. p. 117-120.

LUCAS, C. et al. Computer-assisted text analysis for comparative politics. **Political Analysis**, Oxford, v. 23, n. 2, p. 254-277, Feb. 2015.

MACEDO, D. C.; MATOS, S. N. Extração de conhecimento através da mineração de dados. **Revista de Engenharia e Tecnologia**, Ponta Grossa, v. 2, n. 2, p. 22-30, ago. 2010.

MACHADO, L. et al. A Design Research como método de pesquisa de Administração: Aplicações práticas e lições aprendidas. In: ENCONTRO DA ANPAD, 37., 2013, Rio de Janeiro. **Anais**... Rio de Janeiro: ANPAD, 2013. Disponível em: <http://www.anpad.org.br/admin/pdf/2013_EnANPAD_EPQ748.pdf>. Acesso em: 20 abr. 2016.

MALOUF, R.; MULLEN, T. Taking sides: user classification for informal online political discourse. **Internet Research**, Hackensack, v. 18, n. 2, p. 177-190, Apr. 2008. Disponível em: <http://www.scopus.com/inward/record.url?eid=2-s2.0-41649089293&partnerID=40&md5=9564c357f4ad1e10802fcc1adef2c249>. Acesso em: 20 abr. 2016.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to information retrieval**. Cambridge: Cambridge University, 2008.

MANSON, N. Is operations research really research? **ORiON: The Journal of ORSSA**, Genf, v. 22, n. 2, p. 155-180, 2006.

MANYIKA, J. et al. **Big data**: the next frontier for innovation, competition, and productivity. New York: McKinsey Global Institute, 2011.

MARCH, S. T.; SMITH, G. F. Design and natural science research on information technology. **Decision Support Systems**, Amsterdam, v. 15, n. 4, p. 251-266, Dec. 1995.

MARQUES, P. V.; MELLO, P.; MARTINES FILHO, J. **Mercados futuros e de opções agropecuárias**. Piracicaba: Esalq/USP, 2006. (Didática, n. D-129).

MARTIN, A.; LAKSHMI, T. M.; VENKATESAN, V. P. A business intelligence framework for business performance using data mining techniques. In: INTERNATIONAL CONFERENCE ON EMERGING TRENDS IN SCIENCE, ENGINEERING AND TECHNOLOGY, 1., 2012, Tiruchirappalli. **Proceeedings**… New York: IEEE, 2012. p. 373-380.

MARTIN, N. B.; VEGRO, C. L.; MORICOCHI, L. Custos e rentabilidade de diferentes sistemas de produção de café. **Informações Econômicas**, São Paulo, v. 25, n. 8, p. 131-142, 1995.

MARTINS, C. M. F. **A volatilidade nos preços futuro do café brasileiro e seus principais elementos causadores**. 2005. 154 p. Dissertação (Mestrado em Administração)-Universidade Federal de Lavras, Lavras, 2005.

MARTÍN-VALDIVIA, M.-T. et al. Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. **Expert Systems with Applications**, New York, v. 40, n. 10, p. 3934-3942, Aug. 2013.

MATTMANN, C.; ZITTING, J. **Tika in Action**. Greenwich: Manning Publications, 2011.

MAYRING, P. Qualitative content analysis. **Forum: Qualitative Social Research**, Berlin, v. 1, n. 2, June 2000.

MCCALLUM, A. et al. A machine learning approach to building domain-specific search engines. In: INTERNATIONAL JOINT CONFERENCE ON

ARTIFICIAL INTELLIGENCE, 16., 1999, Stockholm. **Proceedings**… San Francisco: Morgan Kaufmann, 1999. p. 662-667.

MCCALLUM, A.; NIGAM, K. A comparison of event models for naive bayes text classification. In: WORKSHOP ON LEARNING FOR TEXT CATEGORIZATION, 1998, Madison. **Proceedings**… Palto Alto: AAAI, 1998. p. 41-48.

MELÉ, D. Practical wisdom in managerial decision making. **Journal of Management Development**, Bradford, v. 29, n. 7, p. 637-645, 2010. Disponível em: <http://www.scopus.com/inward/record.url?eid=2-s2.0-77955284389&partnerID=40&md5=b747256e3b8a9326068b3481c3c12419>. Acesso em: 20 abr. 2016.

MELO, E. S.; MATTOS, L. B. Análise da volatilidade da base do café arábica para a mesorregião do sul de Minas Gerais. **Revista Economia & Gestão**, Belo Horizonte, v. 12, n. 29, p. 124-140, maio/ago. 2012.

MERCADO, S. **Relatório diário de informações e previsões de mercados interno e externo**. [S.l.]: União dos Produtores de Bionergia, 2014.

MINER, G. **Practical text mining and statistical analysis for non-structured text data applications**. Cambridge, Estados Unidos: Academic Press, 2012.

MITRA, G. et al. **Automated analysis of news to compute market sentiment: its impact on liquidity and trading**. 2011. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=9884AA0F751CF8ECA91D2BDA898FF0CA?doi=10.1.1.372.9679&rep=rep1&type=pdf>. Acesso em: 20 abr. 2016.

MÓL, A. L. R. Value-at-risk da base em operações vendidas de hedge nos contratos futuros de café arábica na BM&F. **Interface**, Natal, v. 5, n. 1, p. 91-108, jan./jun. 2008.

MONTOYO, A.; MARTINEZ-BARCO, P.; BALAHUR, A. Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. **Decision Support Systems**, Amsterdam, v. 53, n. 4, p. 675-679, Nov 2012.

MORAES, R.; VALIATI, J. F.; GAVIAO NETO, W. P. Document-level sentiment classification: An empirical comparison between SVM and ANN.

Expert Systems with Applications, New York, v. 40, n. 2, p. 621-633, Feb. 2013.

MULLER, M.-L. Parts of competitive intelligence: competitor intelligence. **South African Journal of Information Management**, Durbanville, v. 8, n. 1, Mar. 2006.

NAKAGAWA, R.; UCHIDA, H. Herd Behaviour by Japanese Banks after Financial Deregulation. **Economica**, London, v. 78, n. 312, p. 618-636, Oct. 2011.

NAKAZONE, D.; SAES, M. S. M. O agronegócio café do Brasil no mercado internacional. **Revista FAE Business**, Curitiba, n. 9, p. 40-42, set. 2004.
NONAKA, I. A dynamic theory of organizational knowledge creation. **Organization science**, Providence, v. 5, n. 1, p. 14-37, 1994.

NONAKA, I.; VON KROGH, G. Tacit knowledge and knowledge conversion: controversy and advancement in organizational knowledge creation theory. **Organization Science**, Providence, v. 20, n. 3, p. 635-652, May-June 2009.

NUNES, R.; SAES, M. S. M.; BRANDO, J. A. **A volatilidade das cotações de café nas bolsas internacionais**. In: CONGRESSO DA SOBER, 42., 2004, Cuiabá. **Anais**... Cuiabá: Sociedade Brasileira de Economia, Administração e Sociologia Rural, 2004.

OLAREWAJU, A. A. Strategic decision making: A review of literature. **International Business Management**, [S.l.], v. 6, n. 4, p. 552-557, 2012. Disponível em: <http://www.scopus.com/inward/record.url?eid=2-s2.0-84869811376&partnerID=40&md5=3ba055a8314a5453b0d1144b2bbb009c>. Acesso em: 20 abr. 2016.

sheOLIVEIRA, D. H. D. et al. Panorama da cafeicultura de Coffea canephora: perspectivas para Brasil e Vietnã. In: Simpósio de Pesquisa dos Cafés do Brasil, 8., 2013, Salvador. **Anais**... Salvador: Consórcio de Pesquisa do Café.

PAI, M.-Y. et al. Ontology-based SWOT analysis method for electronic word-of-mouth. **Knowledge-Based Systems**, [Amsterdam], v. 50, p. 134-150, Sept. 2013.

PANG, B.; LEE, L. Opinion mining and sentiment analysis. **Foundations and Trends in Information Retrieval**, [S.l.], v. 2, n. 1-2, p. 1-135, 2008.

PAVÃO, A. R. Análise do comportamento da base do Café Arábica: um estudo de caso do município de Alpinópolis – MG. In: CONGRESSO BRASILEIRO DE ECONOMIA, ADMINISTRAÇÃO E SOCIOLOGIA RURAL, 48., 2010, Campo Grande: **Anais...** Brasília: SOBER, 2010.

PELLISSIER, R.; NENZHELELE, T. E. Towards a universal definition of competitive intelligence: original research. **South African Journal of Information Management**, Durbanville, v. 15, n. 2, p. 1-7, 2013a.

_____. Towards a universal competitive intelligence process model. **South Africa Journal of Information Management**, Durbanville, v. 15, n. 2, p. 7, 2013b.

PHADERMROD, B.; CROWDER, R. M.; WILLS, G. B. Developing SWOT analysis from customer satisfaction surveys. In: INTERNATIONAL CONFERENCE ON E-BUSINESS ENGINEERING, 11., 2014, Guangzhou. **Proceedings**… New York: IEEE, 2014. p. 97-104.

PIATESKI, G.; FRAWLEY, W. **Knowledge discovery in databases**. Massachusetts: Massachusetts Institute of Technology, 1991.

PLATT, J. **Sequential minimal optimization**: a fast algorithm for training support vector machines. 1998. Technical Report. Disponível em: <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tr-98-14.pdf>. Acesso em: 20 abr. 2016.

PORTER, M. **Snowball**: a language for stemming algorithms. 2001a. Disponível em: <http://snowball.tartarus.org/texts/introduction.html>. Acesso em: 20 abr. 2016..

PORTER, M. E. **On competition**. Brighton: Harvard Business Press, 2008.

PRAIS, S. J.; WINSTEN, C. B. **Trend estimators and serial correlation**. Cowles Commission discussion paper, Chicago, n. 383, 1954.

QIU, J. et al. Timeline Analysis of Web News Events. In: ZHOU, S.; ZHANG, S.; KARYPIS, G. (Ed.). **Advanced Data Mining and Applications**. Berlin: Springer, 2008. p. 123-134.

QUINLAN, J. R. Induction of decision trees. **Machine learning, Boston,** v. 1, n. 1, p. 81-106, 1986.

_____. **C4. 5**: programs for machine learning. Burlington: Morgan kaufmann, 1993.

RAPP, A.; AGNIHOTRI, R.; BAKER, T. L. Conceptualizing salesperson competitive intelligence: An individual-level perspective. **Journal of Personal Selling & Sales Management,** Münster, v. 31, n. 2, p. 141-155, Mar. 2011.

REGO, B. R.; PAULA, F. O. de. O mercado futuro e a comercialização de café. **Gestão & Conhecimento**, Poços de Caldas, v. 7, n. 1, mar./jun. 2012.

REN, J. et al. Effective Sentiment Analysis of Corporate Financial Reports. In: INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS, 34., 2013, Milan. **Proceedings**… [S.l.: s.n.], 2013.

RENNIE, J. D. et al. Tackling the poor assumptions of naive bayes text classifiers. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 20., 2003, Washington. **Proceedings**… Menlo Park: AAAI, 2003. p. 616-623.

RIBEIRO, K. C. de S.; SOUSA, A. F. de; ROGERS, P. Preços do café no Brasil: variáveis preditivas no mercado à vista e futuro. **Revista de Gestão USP**, São Paulo, v. 13, n. 1, p. 11-30, jan./mar. 2006.

ROMME, A. G. L. Making a difference: organization as design. **Organization science**, Providence, v. 14, n. 5, p. 558-573, Oct. 2003.

SAAYMAN, A. et al. Competitive intelligence: construct exploration, validation and equivalence. **ASLIB Proceedings**, London, v. 60, n. 4, p. 383-411, 2008.

SAMEJIMA, M. et al. SWOT analysis support tool for verification of business strategy. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL CYBERNETICS, 2006, Budapest. **Proceedings**… New York: IEEE, 2006. p. 1-4.

SANTINATO, R.; FERNANDES, A. L. T. Avanços da tecnologia da irrigação na cultura do café. In: SIMPÓSIO DE PESQUISA DOS CAFÉS DO BRASIL, 1., 2000, Poços de Caldas. **Resumos expandidos**… Brasília: Embrapa Café, 2000. Disponível em: <http://www.sbicafe.ufv.br/bitstream/handle/123456789/530/166699_Art15f.pdf?sequence=1&isAllowed=y>. Acesso em: 20 abr. 2016.

SANTOS, C. M. et al. Mercado futuro de café: um estudo de caso. **Registro Contábil**, Recife, v. 3, n. 1, p. 62-84, 2012.

SCHARFSTEIN, D. S.; STEIN, J. C. Herd behavior and investment. **The American Economic Review**, Nashiville, v. 80, n.3, p. 465-479, June 1990.

SCHUMAKER, R. P. et al. Evaluating sentiment in financial news articles. **Decision Support Systems**, Amsterdam, v. 53, n. 3, p. 458-464, June 2012.

SEBASTIANI, F. Machine Learning in Automated Text Categorization. **ACM Computing Surveys**, New York, v. 34, n. 1, p. 1-47, Mar. 2002. Disponível em: <http://www.scopus.com/inward/record.url?eid=2-s2.0-0002442796&partnerID=40&md5=dc1611d885cbf99b11c30addb359b0cb>. Acesso em: 20 abr. 2016.

SELENE XIA, B.; GONG, P. Review of business intelligence through data analysis. **Benchmarking: An International Journal**, Bradford, v. 21, n. 2, p. 300-311, Mar. 2014.

SHELDON, R. **A first course in probability**. New York: Pearson Education India, 2002.

SHEPHERD, D. A.; WILLIAMS, T. A.; PATZELT, H. Thinking about entrepreneurial decision making review and research agenda. **Journal of management**, Stillwater, v. 41, n. 1, p. 11-46, Dec. 2014.

SHI, H.; PENG, C.; XU, M. Z. Business intelligence in construction: a review. **Advanced Materials Research**, [S.l.], v. 594-597, p. 3049-3057, 2012.

SHI, Z. Foundations of intelligence science. **International Journal of Intelligence Science**, [S.l.], v. 1, n. 01, p. 8-16, July 2011.

SIMON, H. A. **The sciences of the artificial**. Cambridge: MIT, 1996.

SINGER, N.; DREHER, F.; LASER, S. Published stock recommendations as institutional investor sentiment in the near-term stock market. **Thünen-Series of Applied Economic Theory**, n. 121, Apr. 2012. Disponível em: <https://www.econstor.eu/bitstream/10419/74658/1/672645130.pdf>. Acesso em: 20 abr. 2016.

SORDI, J. O. de; MEIRELES, M.; SANCHES, C. Design science aplicada às pesquisas em administração: reflexões a partir do recente histórico de

publicações internacionais. **Revista de Administração e Inovação**, São Paulo, v. 8, n. 1, p. 10-36, 2011.

STONE, P. J.; DUNPHY, D. C.; SMITH, M. S. **The general inquirer**: a computer approach to content analysis. Cambridge: MIT, 1966.

SU, J. et al. Discriminative parameter learning for Bayesian networks. In: international conference on Machine learning, 25., 2008, Helsinki. **Proceedings**… New York: ACM, 2008. p. 1016-1023.

SU, Q.; ZHENG, Y.; SWEN, B. Combined approach of web mining and semantic annotation for identifying product features in customer reviews. **Journal of Computational Information Systems**, [S.l.], v. 4, n. 3, p. 1047-1054, June 2008. Disponível em: <http://www.scopus.com/inward/record.url?eid=2-s2.0-48549098441&partnerID=40&md5=9f0626a8f826d091b60d251bbf5af9e6>. Acesso em: 20 abr. 2016.

TAKEDA, H.; VEERKAMP, P.; YOSHIKAWA, H. Modeling design process. **AI magazine**, Menlo Park, v. 11, n. 4, p. 37-48, 1990. ISSN 0738-4602.

TARAPANOFF, K.; GREGOLIN, J. A. R. Inteligência organizacional e competitiva. **Ciência da Informação**, Brasília, v. 31, n. 3, p. 108-109, set./dez. 2002.

_____. Business models, business strategy and innovation. **Long range planning,** London, v. 43, n. 2, p. 172-194, Apr./June 2010.

TEO, T. S.; CHOO, W. Y. Assessing the impact of using the Internet for competitive intelligence. **Information & management**, Amsterdam, v. 39, n. 1, p. 67-83, Oct. 2001.

TETLOCK, P. C. Does public financial news resolve asymmetric information? **Review of Financial Studies**, Cary, v. 23, n. 9, p. 3520-3557, Aug. 2010.

TETLOCK, P. C.; SAAR-TSECHANSKY, M.; MACSKASSY, S. More than words: Quantifying language to measure firms' fundamentals. **The Journal of Finance**, New York,v. 63, n. 3, p. 1437-1467, June 2008.

TRUJILLO, J.; MATÉ, A. Business intelligence 2.0: A general overview. In: _____. (Ed.). **Business Intelligence**. Berlin: Springer, 2012. p. 98-116.

TSYTSARAU, M.; PALPANAS, T. Survey on mining subjective data on the web. **Data Mining and Knowledge Discovery**, Boston, v. 24, n. 3, p. 478-514, 2012.

TURNEY, P. D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: ANNUAL MEETING ON ASSOCIATION FOR COMPUTATIONAL LINGUISTIC, 40., 2002, Philadephia. **Proceedings**… Stroudsburg: Association for Computational Linguistics, 2002. p. 417-424.

VAISHNAVI, V.; KUECHLER, W. **Design research in information systems**. 2004. Disponível em: <http://desrist.org/desrist/content/design-science-research-in-information-systems.pdf>. Acesso: 20 abr. 2016.

VALENTIM, M. L. P. **Métodos qualitativos de pesquisa em Ciência da Informação**. São Paulo: Polis, 2005.

VALKILA, J. Do Fair Trade Pricing Policies Reduce Inequalities in Coffee Production and Trade? **Development Policy Review**, London, v. 32, n. 4, p. 475-493, July 2014.

VAN AKEN, J. E. Management research as a design science: articulating the research products of mode 2 knowledge production in management. **British journal of management**, Chichester, v. 16, n. 1, p. 19-36, Mar. 2005.

VEDDER, R. G. et al. CEO and CIO perspectives on competitive intelligence. **Communications of the ACM**, New York, v. 42, n. 8, p. 108-116, Aug. 1999.

VIVIERS, W.; SAAYMAN, A.; MULLER, M.-L. Enhancing a competitive intelligence culture in South Africa. **International Journal of Social Economics**, Bradford, v. 32, n. 7, p. 576-589, 2005.

WEISS, S. M. et al. **Text mining**: predictive methods for analyzing unstructured information. New York: Springer, 2010.

WIEDEMANN, G. Opening up to big data: Computer-assisted analysis of textual data in social sciences. **Historical Social Research/Historische Sozialforschung**, Mannheim, v. 14, n. 2, p. 332-357, 2013.

WILSON, T. et al. OpinionFinder: A system for subjectivity analysis. In: HLT/EMNLP ON INTERACTIVE DEMONSTRATIONS, 2005, Vancouver.

Proceedings… Stroudsburg: Association for Computational Linguistics, 2005. p. 34-35.

WILSON, T.; WIEBE, J.; HOFFMANN, P. Recognizing contextual polarity in phrase-level sentiment analysis. In: CONFERENCE ON HUMAN LANGUAGE TECHNOLOGY AND EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2005, Sapporo. **Proceedings**… Stroudsburg: Association for Computational Linguistics, 2005. p. 347-354.

WILSON, T.; WIEBE, J.; HWA, R. Just how mad are you? Finding strong and weak opinion clauses. In: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, 19., 2004, San Jose. **Proceedings**… Menlo PArk: AAAI, 1999. p. 761-769.

WIXOM, B. et al. The current state of business intelligence in academia: The arrival of big data. **Communications of the Association for Information Systems**, [S.l.], v. 34, n. 1, p. 1, Jan. 2014.

WOLFRAM, M. S. A. **Modelling the stock market using twitter**. 2010. 65 p. Dissertation (Master of Science Artificial Intelligence)-University of Edinburgh, Edinburgh, 2010.

WU, J.-Y. Computational intelligence-based intelligent business intelligence system: concept and framework. In: INTERNATIONAL CONFERENCE ON COMPUTER AND NETWORK TECHNOLOGY, 2., 2010, Bangkok. **Proceedings**… IEEE, 2010. p. 334-338.

XU, K. et al. Mining comparative opinions from customer reviews for Competitive Intelligence. **Decision Support Systems**, Amsterdam, v. 50, n. 4, p. 743-754, Mar. 2011. Disponível em: <http://www.scopus.com/inward/record.url?eid=2-s2.0-79151480152&partnerID=40&md5=e366595825bfe28e4f026a5f8a49d519>. Acesso em: 20 abr. 2016.

YANG, C.-S.; YE, H.-C. Mining company competitor/collaborator network from online news for competitive intelligence. In: INTERNATIONAL CONFERENCE ON INTELLIGENT TECHNOLOGIES AND ENGINEERING SYSTEMS, 2., 2014, Kaohsiung. **Proceedings**… Cham: Springer, 2014. p. 627-634.

YANG, Y. et al. Text mining and visualization tools - Impressions of emerging capabilities. **World Patent Information**, Oxford, v. 30, n. 4, p. 280-293, Dec. 2008.

YANG, Y.; LIU, X. A re-examination of text categorization methods. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 22., 1999, Berkeley. **Proceedings**… New York: ACM, 1999. p. 42-49.

YU, H.; HATZIVASSILOGLOU, V. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2003, Sapporo. **Proceedings**… Stroudsburg: Association for Computational Linguistics, 2003. p. 129-136.

YU, X. et al. Mining online reviews for predicting sales performance: a case study in the movie domain. **Ieee Transactions on Knowledge and Data Engineering**, New York, v. 24, n. 4, p. 720-734, Apr. 2012.

ZHANG, W.; SKIENA, S. Trading strategies to exploit blog and news sentiment. In: INTERNATIONAL CONFERENCE ON WEBLOGS AND SOCIAL MEDIA, 4., 2010, Washington. **Proceedings**… Palo Alto: AAAI, 2010.

ZHANG, Y. **Forecasting of daily dynamic hedge ratio in agricultural and commodities' futures markets**: evidence from Garch models. 2012. 298 p. Thesis (Doctor of Philosophy in Management)-University of Southampton, Southampton, 2012.

ZIKOPOULOS, P.; EATON, C. **Understanding big data**: analytics for enterprise class hadoop and streaming data. New Jersey: McGraw-Hill Osborne Media, 2011.

ANEXO A – NOTÍCIAS AVALIADAS PELO ESPECIALISTA PARA O REQUISITO AUMENTO DA PRODUÇÃO EM RIVAIS

Descrição: Notícias classificadas como Produção com score maior ou igual a 0.45 e classificadas como positivas ou negativas para oferta.

26/05/2016	The Exchange (press release) (blog)	Effort to revamp coffee production in Tanzania underway	http://exchange.co.tz/effort-to-revamp-coffee-production-in-tanzania-underway/
17/05/2016	Andina - Agencia Peruana de Noticias	Peru coffee sector to grow 15%, shipments to total US$700 million in ...	http://www.andina.com.pe/ingles/noticia-peru-coffee-sector-to-grow-15-shipments-to-total-700-million-in-2016-612998.aspx
17/05/2016	Vending Times	Colombian Coffee Production Increased 13% In April, But El Niño ...	http://www.vendingtimes.com/ME2/dirmod.asp?nm=Vending Features&type=Publishing&mod=Publications::Article&tier=4&id=6235546231C848D8A18EFC3EF13765FB
15/05/2016	Coastweek	Tanzania to phase out khat planting in favor of coffee farming	http://www.coastweek.com/3920-Tanzania-to-phase-out-khat-planting-in-favor-of-coffee-farming.htm
12/05/2016	Andina - Agencia Peruana de Noticias	Peruvian green coffee re-enters Brazilian market after almost one year	http://www.andina.com.pe/ingles/noticia-peruvian-green-coffee-reenters-brazilian-market-after-almost-one-year-612283.aspx
10/05/2016	New Vision	Uganda Coffee Development Authority (UCDA)	http://www.newvision.co.ug/new_vision/news/1424235/uganda-coffee-development-authority-ucda
09/05/2016	Andina - Agencia Peruana de Noticias	Peru: Coffee, Amazonas top export product	http://www.andina.com.pe/Ingles/noticia-peru-coffee-amazonas-top-export-product-611779.aspx
08/05/2016	RFI	Tanzania: High hopes for Magufuli's first budget	http://en.rfi.fr/africa/20160508-tanzania-magufuli-first-budget-minister-speaks
02/05/2016	Reuters	Honduras coffee exports rise 3.2 pct in April	http://www.reuters.com/article/honduras-coffee-idUSL2N18000J

Decréscimo de Produção

25/05/2016	South China Morning Post	As rains begin in Vietnam, coffee farmers count cost of drought ...	http://www.scmp.com/news/asia/southeast-asia/article/1954415/rains-begin-top-coffee-producer-vietnam-farmers-count-cost
25/05/2016	Reuters	As Vietnam's rains begin, coffee farmers eye drought damage	http://www.reuters.com/article/us-vietnam-coffee-drought-idUSKCN0YG0KB
19/05/2016	Agrimoney.com	Colombian coffee output to dip for first time in five years	http://www.agrimoney.com/news/colombian-coffee-output-to-dip-for-first-time-in-five-years--9567.html
17/05/2016	Times of India	India coffee output to fall to lowest in two decades	http://timesofindia.indiatimes.com/business/india-business/India-coffee-output-to-fall-to-lowest-in-two-decades/articleshow/52308693.cms
10/05/2016	Bloomberg	Indonesia Coffee Crop May Tumble Most in 5 Years on Drought	http://www.bloomberg.com/news/articles/2016-05-09/indonesia-coffee-crop-may-tumble-most-in-five-years-on-drought
10/05/2016	Indonesia Investments (press release)	Indonesia's Coffee Output Down on El Nino and La Nina'	http://www.indonesia-investments.com/news/todays-headlines/indonesia-s-coffee-output-down-on-el-nino-and-la-nina/item6800
09/05/2016	Bangkok Post	Indonesian coffee crop down 10% as drought bites	http://www.bangkokpost.com/news/asia/966913/indonesian-coffee-crop-down-10-as-drought-bites
09/05/2016	Andina - Agencia Peruana de Noticias	Peru: Vraem to become world's largest cocoa supplier in 10 yrs	http://www.andina.com.pe/ingles/noticia-peru-vraem-to-become-worldâ€™s-largest-cocoa-supplier-in-10-yrs-611770.aspx
05/05/2016	VietNamNet Bridge	Long drought hits crops in Vietnam	http://english.vietnamnet.vn/fms/environment/155999/long-drought-hits-crops-in-vietnam.html
04/05/2016	Sudan Tribune	Ethiopia to invest billions on green economy projects	http://www.sudantribune.com/spip.php?article58863
03/05/2016	VietNamNet Bridge	Worst drought in 30 years strikes Vietnam coffee	http://english.vietnamnet.vn/fms/business/155882/worst-drought-in-30-years-strikes-vietnam-coffee.html
01/05/2016	The Rakyat Post	Big fall in Indonesia's export of robusta coffee	http://www.therakyatpost.com/business/2016/05/02/big-fall-in-indonesias-export-of-robusta-coffee/

ANEXO B - AVALIAÇÃO DAS INFORMAÇÕES DO PROTÓTIPO PARA ANÁLISE DE REQUISITOS DE INTELIGÊNCIA

O protótipo apresenta notícias relacionadas a um requisito de inteligência para a Cafeicultura. A partir da análise das notícias responda as questões.

Requisito Avaliado: Produção em Rivais

1) As fontes das notícias são relevantes e frequentemente consultadas para análise?

Sim, o protótipo selecionou notícias de diversas fontes comumente utilizadas pela equipe do Bureau de Inteligência Competitiva do Café, como Reuters, Bloomberg, Agrimoney, que são reconhecidas mundialmente. Ele também foi capaz de selecionar notícias de websites locais, mas que são relevantes para nossas análises, como VietNamNet e Times of India. Surgiram também novas fontes, que poderão ser avaliadas quanto a sua relevância.

2) As notícias apresentadas são relevantes para o requisito de IC estudado?

Com relação ao aumento ou redução da oferta mundial de grãos, as notícias são bastante relevantes. Foram identificadas diversas notícias que tratam de ameaças decorrentes de fatores climáticos e de previsões de aumento devidas, também, a fatores climáticos ou iniciativas governamentais.

3) O sistema contribui para identificar oportunidades, ameaças, forças e fraquezas?

O sistema contribui principalmente para a identificação de “oportunidades” e “ameaças”, embora ainda ocorram algumas imprecisões, como a classificação de algo que seria uma ameaça a um concorrente como ameaça ao Brasil. Quanto as forças e fraquezas, para o período analisada o protótipo identificou apenas uma notícia, sendo que esta classificação estava errada. De modo geral, o recurso é promissor e será útil, mas precisa ser melhor calibrado.

4) O sistema contribui para identificar evidencias que impactam oferta e demanda?

Sim, o sistema foi capaz de identificar inúmeras evidencias que poderão impactar a oferta. Nesse tipo de informação, a comprovação do impacto só pode ser verificada meses ou anos após a notícia. Por outro lado, algumas notícias de "baixa relevância" para a oferta acabaram sendo selecionadas também.

5) Pela análise das notícias é possível acrescentar informação competitiva não capturada sem o sistema? Se sim, qual?

Em comparação com a busca manual de notícias realizada pelo Bureau no mês de maio, o sistema foi capaz de identificar notícias relevantes sobre a Tanzânia.

6) O artefato contribui para o requisito de IC?

Sim, o artefato se mostrou útil para a identificação de informações relevantes para a produção de café no mundo.

Informação do Participante:

A informação coletada é usada apenas para pesquisa. Todas as informações serão mantidas confidenciais.

Empresa: Centro de Inteligência em Mercados

Cargo: Coordenador de projeto

Por quanto tempo está neste cargo?: 6 anos

Descreva sua qualificação profissional e experiência:

Especialista em análises de inteligência competitiva para o agronegócio café.

Descreva sua formação acadêmica:

Doutorando em Administração e mestre em Administração pela UFLA.

ANEXO C – NOTÍCIAS AVALIADAS PELO ESPECIALISTA PARA O REQUISITO DESENVOLVIMENTO DA INDÚSTRIA

Descrição: Notícias classificadas como Industria com score maior ou igual a 0.45 e classificadas como positivas ou negativas para demanda.

31/05/2016	Vending Times	JM Smucker Lowers Folgers, Dunkin' Donuts Coffee Prices 6%	http://www.vendingtimes.com/ME2/dirmod.asp?sid=EB79A487112B48A296B38C81345C8C7F&nm=Vending Features&type=Publishing&mod=Publications::Article&mid=8F3A7027421841978F18BE895F87F791&tier=4&id=D22A7855C38649C694908AC0567D6875
31/05/2016	GlobeNewswire (press release)	Marley Coffee Issues Shareholder Letter to Discuss Anticipated ...	https://globenewswire.com/news-release/2016/05/31/844623/0/en/Marley-Coffee-Issues-Shareholder-Letter-to-Discuss-Anticipated-Results-of-Operations-For-Its-2017-First-Quarter-Three-Months-Ended-April-30-2016.html
26/05/2016	International Business Times	Starbucks Corp. (SBUX) Opening Coffee Roastery In China, Moving ...	http://www.ibtimes.com/starbucks-corp-sbux-opening-coffee-roastery-china-moving-outside-us-first-time-2374411
25/05/2016	Bloomberg	Nestle Downplays Risk of Brexit to World's Largest Food Company	http://www.bloomberg.com/news/articles/2016-05-25/nestle-downplays-risk-of-brexit-to-world-s-largest-food-company
25/05/2016	AgraNet (subscription)	Italy - Lavazza 2015 profit jumps after sale of Keurig Green Mountain ...	https://www.agra-net.com/agra/international-coffee-report/analysis/company/italy---lavazza-2015-profit-jumps-after-sale-of-keurig-green-mountain-stake--1.htm
24/05/2016	Chicago Tribune	Coffee cost cuts: Lower prices for Folgers, Dunkin' Donuts store ...	http://www.chicagotribune.com/business/ct-coffee-prices-lower-0525-biz-20160524-story.html
24/05/2016	Fortune	Bags of Folgers and Dunkin' Donuts Coffee Are Getting Even Cheaper	http://fortune.com/2016/05/24/folgers-dunkin-donuts-coffee/
24/05/2016	Yahoo News	Folgers coffee maker JM Smucker to cut prices in US	http://finance.yahoo.com/news/folgers-coffee-maker-j-m-131635390.html
24/05/2016	Daily News & Analysis	Tata Global Beverages net at Rs 326 crore, up 32%	http://www.dnaindia.com/money/report-tata-global-beverages-net-at-rs-326-crore-up-32-2216328
24/05/2016	Fusion	Get ready for the Keurig of weed, environment	http://fusion.net/story/305750/single-use-marijuana-vape-like-keurig/
21/05/2016	PR Newswire (press release)	Lavazza to launch Lavazza Â¡Tierra! and Grand Hotel at the 2016 ...	http://www.prnewswire.com/news-releases/lavazza-to-launch-lavazza-tierra-and-grand-hotel-at-the-2016-national-restaurant-association-show-300272780.html
20/05/2016	Allentown Morning Call	Kraft Heinz moving local Tassimo coffee production to Canada	http://www.mcall.com/news/breaking/mc-kraft-workers-allentown-taa-petition-20160520-story.html
19/05/2016	Data Center Knowledge	Tata Communications Sells 17 Data Centers for $633M	http://www.datacenterknowledge.com/archives/2016/05/19/tata-communications-sells-17-data-centers-633m/

19/05/2016	Allentown Morning Call	Kraft Heinz to close local plant this summer, reimburse state $200K	http://www.mcall.com/news/breaking/mc-pa-clawback-money-kraft-heinz-20160518-story.html
19/05/2016	Seeking Alpha	Krispy Kreme Merger-Arb Not Worth The Risk	http://seekingalpha.com/article/3976452-krispy-kreme-merger-arb-worth-risk
19/05/2016	Kent Online	Costa Coffee to open shop on Sheppey	http://www.kentonline.co.uk/sheerness/news/neats-costa-96171/
17/05/2016	Times of India	Tata Coffee shares down 3% as Q4 net dips	http://timesofindia.indiatimes.com/city/mumbai/Tata-Coffee-shares-down-3-as-Q4-net-dips/articleshow/52310491.cms
17/05/2016	Economic Times	Tata Coffee shares down 3% as Q4 net dips	http://articles.economictimes.indiatimes.com/2016-05-17/news/73152040_1_consolidated-net-profit-tata-coffee-total-income
17/05/2016	Reuters	Kraft, Starbucks defeat appeal of coffee pod settlement	http://www.reuters.com/article/starbucks-tassimo-idUSL2N18E0AC
17/05/2016	BakeryAndSnacks.com	TSW Foods to bring Krispy Kreme products to convenience stores	http://www.bakeryandsnacks.com/Manufacturers/TSW-Foods-to-bring-Krispy-Kreme-products-to-convenience-stores
16/05/2016	Moneycontrol.com	Petronet LNG, JK Tyres, Tata Coffee in limelight post Q4 nos	http://www.moneycontrol.com/news/stocks-to-watch/petronet-lng-jk-tyres-tata-coffeelimelight-post-q4-nos_6670381.html
13/05/2016	Agrimoney.com	JAB Holdings continues coffee consolidation with Krispy Kreme ...	http://www.agrimoney.com/news/jab-holdings-continues-coffee-consolidation-with-krispy-kreme-acquisition--9556.html
12/05/2016	FoodBev.com	Lavazza in 'important' three-year link-up with Italian airline Alitalia	http://www.foodbev.com/news/lavazza-in-important-three-year-link-up-with-italian-airline-alitalia/
11/05/2016	News Tribune	Starbucks Corporation (NASDAQ:SBUX): Stock Target from Analysts	http://www.greenvilletribune.com/starbucks-corporation-nasdaqsbux-stock-target-from-analysts/
11/05/2016	Vending Times	Keurig Owner JAB Continues Building Coffee Empire With Krispy ...	http://www.vendingtimes.com/ME2/dirmod.asp?sid=EB79A487112B48A296B38C81345C8C7F&nm=Vending Features&type=Publishing&mod=Publications::Article&mid=8F3A7027421841978F18BE895F87F791&tier=4&id=87B1166578BE46248D260F29F3E4EE58
11/05/2016	The FINANCIAL	New partnership between Alitalia and Lavazza	http://www.finchannel.com/index.php/tourism-and-travel/57217-new-partnership-between-alitalia-and-lavazza
10/05/2016	Christian Science Monitor	Krispy Kreme will be acquired by JAB Holding in $1.35-billion deal	http://www.csmonitor.com/Business/The-Bite/2016/0509/Krispy-Kreme-will-be-acquired-by-JAB-Holding-in-1.35-billion-deal
10/05/2016	Channel News Asia	JAB to take Krispy Kreme private for US$1.35 billion	http://www.channelnewsasia.com/news/business/international/jab-to-take-krispy-kreme/2770082.html
10/05/2016	Seeking Alpha	Krispy Kreme Taken Private, Will Cheesecake Factory Be Next?	http://seekingalpha.com/article/3973554-krispy-kreme-taken-private-will-cheesecake-factory-next
10/05/2016	FOODBEAST	The Owners Of Keurig Just Bought Krispy Kreme, And F*ck ...	http://www.foodbeast.com/news/keurig-krispy/

10/05/2016	Market Realist	Did Kraft Heinz's Revenue Meet Expectations in Fiscal 1Q16?	http://marketrealist.com/2016/05/kraft-heinzs-revenue-meet-expectations-fiscal-1q16/
10/05/2016	China Post	Krispy Kreme shares get jolt from coffee chain JAB Beech takeover	http://www.chinapost.com.tw/business/company-focus/2016/05/11/465771/Krispy-Kreme.htm
10/05/2016	Benzinga	Pooki's Mahi Shipping Keurig 2.0 Compatible 100% Kona Coffee K ...	http://www.benzinga.com/pressreleases/16/05/p7963506/pookis-mahi-shipping-keurig-2-0-compatible-100-kona-coffee-k-cups
10/05/2016	NOLA.com	Krispy Kreme to be sold to coffee company for $1.35 billion	http://www.nola.com/business/index.ssf/2016/05/krispy_kreme_sold_reimann.html
10/05/2016	Reuters	JAB to take Krispy Kreme private for $1.35 billion	http://www.reuters.com/article/us-krispykreme-m-a-jab-idUSKCN0Y01EM
09/05/2016	Reuters	Germany's JAB to take Krispy Kreme private for $1.35 billion	http://www.reuters.com/article/uk-krispykreme-m-a-jab-idUSKCN0Y01EM
09/05/2016	Reuters	UPDATE 3-Germany's JAB to take Krispy Kreme private for $1.35 bln	http://www.reuters.com/article/krispykreme-ma-jab-idUSL3N1864H8
09/05/2016	Times of India	Krispy Kreme being taking private in $1.35 billion deal	http://timesofindia.indiatimes.com/business/international-business/Krispy-Kreme-being-taking-private-in-1-35-billion-deal/articleshow/52193009.cms
09/05/2016	Channel News Asia	Germany's JAB to take Krispy Kreme private for US$1.35 billion	http://www.channelnewsasia.com/news/business/international/germany-s-jab-to-take-kri/2770082.html
09/05/2016	share market updates (press release)	Recent Movements of Stocks: Kraft Heinz Co (KHC), Pilgrim's Pride ...	http://sharemarketupdates.com/recent-movements-of-stocks-kraft-heinz-co-khc-pilgrims-pride-corporation-ppc-pepsico-inc-pep/
09/05/2016	BBC News	Krispy Kreme bought by Kenco coffee owner	http://www.bbc.co.uk/news/business-36248176
09/05/2016	The Japan Times	Sweet $1.35 billion takeover offer gives Krispy Kreme shares ...	http://www.japantimes.co.jp/news/2016/05/10/business/sweet-1-35-billion-takeover-offer-gives-krispy-kreme-shares-caffeine-spike/
09/05/2016	New York's PIX11 / WPIX-TV	Krispy Kreme bought for $1.35 billion	http://pix11.com/2016/05/09/krispy-kreme-bought-for-1-35-billion/
09/05/2016	12NewsNow.Com	Krispy Kreme shares soar on sweet takeover offer	http://www.12newsnow.com/story/31923796/krispy-kreme-shares-soar-on-sweet-takeover-offer
09/05/2016	Wall Street Journal	Krispy Kreme to Be Acquired by Keurig Owner JAB for $1.35 Billion	http://www.wsj.com/articles/krispy-kreme-to-be-acquired-by-jab-for-1-35-billion-1462798137
09/05/2016	Daily Mail	Krispy Kreme set to be taken over by coffee giant in deal worth $1.35 ...	http://www.dailymail.co.uk/news/article-3581003/Krispy-Kreme-taking-private-1-35B-deal.html
09/05/2016	Christian Science Monitor	Krispy Kreme bought by coffee-loving JAB for $1.35 billion	http://www.csmonitor.com/Business/The-Bite/2016/0509/Krispy-Kreme-bought-by-coffee-loving-JAB-for-1.35-billion

09/05/2016	TwinCities.com-Pioneer Press	Coffee & doughnuts: Caribou Coffee owner buying Krispy Kreme	http://www.twincities.com/2016/05/09/doughnuts-coffee-caribou-coffee-owner-buying-krispy-kreme/
09/05/2016	Crow River Media	Another Caribou Coffee in the works	http://www.crowrivermedia.com/hutchinsonleader/news/business/another-caribou-coffee-in-the-works/article_e5767bbc-60bc-5f48-b569-de614a690cf2.html
09/05/2016	Business Insider	The investor that's buying Krispy Kreme is building a coffee and ...	http://www.businessinsider.com/jab-holding-building-a-coffee-and-bagel-empire-2016-5
09/05/2016	eNCA	Krispy Kreme bought by JAB, owner of big coffee brands	https://www.enca.com/money/krispy-kreme-bought-by-jab-owner-of-big-coffee-brands
09/05/2016	The Providence Journal	Coffee giant consuming Krispy Kreme in $1.35-billion deal	http://www.providencejournal.com/article/20160509/NEWS/160509340
09/05/2016	Seeking Alpha	Why Hasn't Coca-Cola Bought Monster Yet?	http://seekingalpha.com/article/3973013-coca-cola-bought-monster-yet
09/05/2016	Yahoo News	Germany's JAB to take Krispy Kreme private for $1.35 billion	http://finance.yahoo.com/news/germanys-jab-krispy-kreme-private-135817952.html
09/05/2016	Daily Mail	Krispy Kreme bought by JAB, owner of big coffee brands	http://www.dailymail.co.uk/wires/afp/article-3581452/Krispy-Kreme-bought-JAB-owner-big-coffee-brands.html
09/05/2016	DIGITALLOOK	Douwe Egberts owner gobbles up Krispy Kreme to add to coffee stable	http://www.digitallook.com/news/international-economic/douwe-egberts-owner-gobbles-up-krispy-kreme-to-add-to-coffee-stable--1156797.html
09/05/2016	Powder Bulk Solids	JAB Holdings Acquires Krispy Kreme to Challenge Nestle	http://www.powderbulksolids.com/news/JAB-Holdings-Acquires-Krispy-Kreme-to-Challenge-Nestle-05-09-2016
09/05/2016	The Guardian	Donut with your coffee? Krispy Kreme sold to Keurig and Stumptown ...	https://www.theguardian.com/business/2016/may/09/krispy-kreme-doughnuts-sold-reimann-family-germany
09/05/2016	Channel News Asia	Germany's JAB to take Krispy Kreme private for US$1.35 billion	http://www.channelnewsasia.com/news/business/germany-s-jab-to-take-kri/2770082.html
09/05/2016	Law360 (subscription)	Investment Firm JAB Inks $1.35B Krispy Kreme Buyout	http://www.law360.com/articles/793861/investment-firm-jab-inks-1-35b-krispy-kreme-buyout
09/05/2016	The Consumerist	Keurig Parent Company Wants Doughnuts With That Coffee, Buys ...	https://consumerist.com/2016/05/09/keurig-parent-company-wants-doughnuts-with-that-coffee-buys-krispy-kreme/
09/05/2016	WUNC	Sweet Deal? Krispy Kreme Is Being Acquired By German ...	http://wunc.org/post/sweet-deal-krispy-kreme-being-acquired-german-conglomerate
08/05/2016	Franklin Independent	How Analysts Rated Keurig Green Mountain Inc (NASDAQ:GMCR ...	http://www.franklinindependent.com/how-analysts-rated-keurig-green-mountain-inc-nasdaqgmcr-last-week/
07/05/2016	Franklin Independent	How Many Keurig Green Mountain Inc (NASDAQ:GMCR)'s Analysts ...	http://www.franklinindependent.com/how-many-keurig-green-mountain-inc-nasdaqgmcrs-analysts-are-bearish/
06/05/2016	Fox News	Krispy Kreme launches edible coffee squares	http://www.foxnews.com/leisure/2016/05/06/krispy-kreme-launches-edible-coffee-squares/

06/05/2016	The Point Review	Noticeable Earnings Watch: Kraft Heinz Co (NASDAQ:KHC)	http://www.thepointreview.com/noticeable-earnings-watch-kraft-heinz-co-nasdaqkhc/
04/05/2016	International Business Times, India Edition	Tata Group published 3500 patents in two years	http://www.ibtimes.co.in/tata-group-published-3500-patents-two-years-677420
03/05/2016	Fox News Latino	Starbucks sued in US for putting 'too much ice' in cold drinks	http://latino.foxnews.com/latino/lifestyle/2016/05/03/starbucks-sued-in-us-for-putting-too-much-ice-in-cold-drinks/
03/05/2016	Quartz	Soon you'll be able to get your tortillas and pet food from Keurig-like ...	http://qz.com/674609/pods-are-turning-food-into-the-most-profitable-business-since-software/
03/05/2016	China Post	Woman sues Starbucks over ice in cold drinks	http://www.chinapost.com.tw/business/company-focus/2016/05/04/465062/Woman-sues.htm
03/05/2016	InterAksyon	US woman sues Starbucks for $5 million over false advertising ...	http://interaksyon.com/article/127273/us-woman-sues-starbucks-for-5-million-over-false-advertising-consumer-fraud
02/05/2016	Mirror.co.uk	Woman sues Starbucks for Â£3.4million claiming iced drinks have too ...	http://www.mirror.co.uk/news/world-news/woman-sues-starbucks-34million-claiming-7884091
02/05/2016	Franklin Independent	Are Analysts Bearish Dunkin Brands Group Inc (NASDAQ:DNKN ...	http://www.franklinindependent.com/are-analysts-bearish-dunkin-brands-group-inc-nasdaqdnkn-after-last-week/

ANEXO D - AVALIAÇÃO DAS INFORMAÇÕES DO PROTÓTIPO PARA ANÁLISE DE REQUISITOS DE INTELIGÊNCIA

O protótipo apresenta notícias relacionadas a um requisito de inteligência para a Cafeicultura. A partir da análise das notícias responda as questões.

Requisito Avaliado: Desenvolvimento da Indústria

1) As fontes das notícias são relevantes e frequentemente consultadas para análise?

Sim. O protótipo selecionou notícias das principais fontes consultadas, como Reuters, Vending Times e Seeking Alpha.

2) As notícias apresentadas são relevantes para o requisito de IC estudado?

Sim. O protótipo identificou notícias de grande relevância para o requisito, no entanto o resultado apresentou pouca variedade de assuntos. Para o mês de maio, a maioria das notícias selecionadas tratavam de apenas três fatos específicos. Notícias de menor visibilidade, mas relevantes para o requisito podem ter ficado de fora.

3) O sistema contribui para identificar oportunidades, ameaças, forças e fraquezas?

O sistema contribui de maneira razoável para a identificação de oportunidades, função que é prejudicada por algumas notícias classificadas de maneira inadequada. Também houve sobreposição de notícias classificadas tanto como oportunidades quanto forças. Os critérios para forças precisam aperfeiçoados, talvez com a ampliação da amostra classificada manualmente para o aprendizado do algoritmo. Nos atributos fraquezas e ameaças não foram identificadas notícias.

4) O sistema contribui para identificar evidencias que impactam oferta e demanda?

Sim. A partir da leitura das notícias selecionadas pelo sistema é possível elaborar um julgamento sobre a oferta e demanda de café industrializado.

5) Pela análise das notícias é possível acrescentar informação competitiva não capturada sem o sistema? Se sim, qual?

Para a amostra referente ao mês de maio, o sistema não acrescentou novas informações ao que já havia sido identificado manualmente. Talvez o viés em relação à apenas três temas principais tenha contribuído para isso. No entanto, na parte operacional o sistema se mostra útil para agilizar o processo de busca e análise realizado pelos analistas da equipe.

6) O artefato contribui para o requisito de IC?

Sim, o artefato se mostrou útil para a identificação de informações relevantes e também poderá contribuir para a parte operacional do trabalho. Algumas limitações poderão ser solucionadas no futuro, tornando o artefato mais eficiente.

Informação do Participante:

A informação coletada é usada apenas para pesquisa. Todas as informações serão mantidas confidenciais.

Empresa: Centro de Inteligência em Mercados

Cargo: Coordenador de projeto

Por quanto tempo está neste cargo?: 6 anos

Descreva sua qualificação profissional e experiência:

Especialista em análises de inteligência competitiva para o agronegócio café.

Descreva sua formação acadêmica:

Doutorando em Administração e mestre em Administração pela UFLA.

ANEXO E – DADOS MENSAIS

Datas	Média	Fechamento	Produção	Indústria	Oferta (+)
jan/11	235,96	244,80	4	5	0
fev/11	258,91	271,70	11	17	1
mar/11	272,07	264,15	8	13	0
abr/11	281,66	299,35	7	5	0
mai/11	275,07	264,60	5	15	1
jun/11	258,01	265,35	17	2	0
jul/11	251,55	239,55	11	4	0
ago/11	256,43	289,10	13	7	1
set/11	259,98	228,90	17	5	1
out/11	232,80	226,95	20	13	2
nov/11	230,67	233,80	21	5	0
dez/11	223,30	226,85	12	4	0
jan/12	223,25	215,05	16	10	2
fev/12	208,87	203,15	16	13	0
mar/12	186,45	182,45	15	17	1
abr/12	178,63	177,95	15	8	0
mai/12	173,63	160,65	23	15	1
jun/12	156,80	170,10	21	19	2
jul/12	180,03	174,40	17	12	1
ago/12	165,74	164,55	21	12	0
set/12	171,70	228,90	17	10	2
out/12	164,74	154,65	24	13	2
nov/12	147,60	142,10	28	17	0
dez/12	141,37	143,80	28	10	3
jan/13	150,03	146,95	23	4	3
fev/13	141,43	142,65	20	9	2
mar/13	138,88	137,15	21	8	0
abr/13	137,19	134,95	18	12	0
mai/13	135,99	127,05	24	13	2
jun/13	123,36	120,00	23	20	2
jul/13	122,82	118,60	25	12	3

ago/13	117,61	112,10	13	7	0
set/13	114,38	113,70	37	9	2
out/13	112,79	105,40	32	9	2
nov/13	105,60	110,25	29	13	2
dez/13	111,69	110,70	30	7	1
jan/14	117,62	125,20	26	12	1
fev/14	154,18	179,80	51	30	2
mar/14	188,49	177,90	40	20	1
abr/14	197,02	203,05	37	16	1
mai/14	186,90	177,50	58	13	0
jun/14	171,60	173,00	31	13	1
jul/14	171,14	195,05	33	19	1
ago/14	187,63	195,75	32	29	0
set/14	186,19	193,35	34	19	3
out/14	205,05	188,00	47	22	4
nov/14	188,18	186,65	37	17	7
dez/14	174,98	166,60	39	34	0
jan/15	168,59	161,90	41	16	4
fev/15	154,90	136,75	37	23	2
mar/15	134,70	132,90	35	18	6
abr/15	138,42	136,55	36	25	3
mai/15	131,46	126,15	61	18	4
jun/15	131,80	130,65	44	14	5
jul/15	124,37	120,40	36	17	5
ago/15	127,14	120,55	45	28	6
set/15	116,64	121,35	49	16	4
out/15	125,13	120,95	45	11	6
nov/15	118,14	116,90	50	15	4
dez/15	120,40	126,70	39	34	4