XVIII SEMEAD
Seminários em Administração
novembro de 2015
ISSN 2177-3866

Avaliação de Métodos de Classificação Textual para Apoio a Análise de Conteúdo Aplicada a Gestão da Informação no Mercado de Café

PAULO DE OLIVEIRA LIMA JÚNIOR
Centro Federal de Educação Tecnológica de Minas Gerais - CEFETMG
plimajunior@gmail.com

LUIZ GONZAGA DE CASTRO JUNIOR
Universidade federal de lavras
lgcastro@ufla.br

ANDRE LUIZ ZAMBALDE
UFLA
zambaufla@gmail.com

Área Temática: Tecnologia da Informação

Avaliação de Métodos de Classificação Textual para Apoio a Análise de Conteúdo Aplicada a Gestão da Informação no Mercado de Café

Resumo: O artigo verifica a viabilidade de classificação textual supervisionada utilizando métodos de aprendizado de máquina para promover a Análise de Conteúdo de notícias em um sistema de apoio a tomada de decisões na cafeicultura. Para tal é desenvolvido um artefato que possibilita comparação empírica dos métodos Naive Bayes, Árvores de Decisão e Support Vector Machines (SVM) para classificar notícias coletadas da web de acordo com categorias pré-definidas da cadeia produtiva do café. Os testes mostram melhor desempenho e viabilidade dos classificadores Bayesianos para o contexto estudado e contribui para redução de recursos necessários no processo de análise.

Abstract: The article verifies the feasibility of supervised text classification using machine learning to promote content analysis of news from coffee market. For this purpose we develop a device that enables empirical comparison of methods Naive Bayes, Decision Trees and Support Vector Machines (SVM) to classify news collected from web in predefined categories based on coffee supply chain in a real information system to support decisions on coffee agribusiness. Tests show better performance and viability of Bayesian classifiers to the context studied and it contributes to the reduction of resources in the analysis process.

Palavras-chave: Categorização Textual, Aprendizado de Máquina, Análise de Conteúdo.

1. INTRODUÇÃO

A variação na oferta e demanda mundial de café causa impactos em diferentes setores da sua cadeia produtiva. O risco é elevado principalmente pelo comportamento do preço do café, influenciado por fatores de mercado, climáticos e macroeconômicos como taxa de juros, cambio e política monetária.

Este ambiente dinâmico exige agilidade dos agentes no delineamento de cenários para previsões de mercado, e sistemas de informação capazes de lidar não só com dados quantitativos para construção de modelos matemáticos, mas também com dados qualitativos para interpretação de eventos, comportamentos dos agentes e tendências.

Na web, o crescimento exponencial de notícias na mídia especializada em cafeicultura, é um repositório mundial vasto e dinâmico de dados textuais com informação relevante, desde a produção até preferencias de consumo. Porém a interpretação de notícias foge da objetividade e quantificação, e suscita metodologias para análise qualitativa.

Paralelamente, é crescente a abordagem qualitativa em estudos na Administração com aplicação de diferentes técnicas (Mayring, 2000). Entre elas, a Análise de Conteúdo é apropriada para tratamento de texto de maneira sistemática e baseada em teoria. É aplicada a dados coletados em fontes de acordo com o contexto da pesquisa, como entrevistas, obras literárias, artigos de jornais e revistas entre outros. E em diferentes formas de documentação de material, não somente textual, mas imagem, filme, áudio e outras relevantes para a análise.

Quando se trata de notícias publicadas na web, nem sempre os dados estão organizados adequadamente, a disponibilidade de informações é difusa, não estruturada, com ruídos e exige recursos que dificultam sua utilização. Diante destes desafios, pesquisas têm impulsionado o avanço de técnicas computacionais de recuperação, processamento e análise

que possibilitam o uso da informação textual por sistemas para tomada de decisões em diferentes contextos: mercado de ações (Lee, Wu, & Chen, 2012), (Li, Liang, Li, Wang, & Wu, 2009), (Schumaker, Zhang, Huang, & Chen, 2012), monitoramento de epidemias (Collier et al., 2008), lançamento e revisão de produtos (Q. Su, Zheng, & Swen, 2008), (Xu, Liao, Li, & Song, 2011), (Yu, Liu, Huang, & An, 2012) e política (Malouf & Mullen, 2008) (Junqué De Fortuny, De Smedt, Martens, & Daelemans, 2012). No mercado de commodities os trabalhos tem foco em petróleo e ouro, conforme apresentado em (Feuerriegel & Neumann, 2013).

Um caso de análise qualitativa para a cafeicultura brasileira é o Bureau de Inteligência Competitiva do Café, projeto do Centro de Inteligência em Mercados da Universidade Federal de Lavras. A partir da Análise de Conteúdo, especialistas pesquisam notícias publicadas na web por fontes especializadas no mercado de café, classificam em categorias temáticas pré-definidas de acordo com setores da cadeia produtiva e produzem um relatório que permite o delineamento de cenários para tomada de decisões e criação de Inteligência Competitiva.

A Análise de Conteúdo é contínua para produção mensal de relatórios pelo Bureau, porém a coleta e classificação são restritas a capacidade de busca e leitura dos especialistas o que consome recursos humanos e tempo. Assim a análise é comprometida à medida que aumentam as fontes de notícias e entidades monitoradas para Inteligência Competitiva.

Diante da necessidade de agilidade para análise de dados qualitativos em notícias, o objetivo deste trabalho é verificar a viabilidade de técnicas computacionais de processamento textual para promover o sistema de gerenciamento de informações e a Análise de Conteúdo do Bureau. A técnica adequada ao cenário descrito é a classificação textual supervisionada através de aprendizado de máquina, usada para categorizar documentos de acordo com seu conteúdo, em classes pré-definidas.

Para tal é desenvolvido um sistema que possibilita comparação empírica entre diferentes versões de métodos de aprendizado de máquina utilizando a base existente no Bureau para treinamento e teste. Além da comparação o sistema é capaz de coletar e classificar novas notícias.

O experimento mostra que os métodos Bayesianos tem melhor desempenho para classificação de notícias extraídas da *web* em comparação aos demais quando confrontadas com a classificação realizada pelos especialistas. Os resultados apontam a viabilidade do uso de classificadores pelo Bureau para automatização da coleta e classificação, ganho de escala e redução de recursos humanos, bem como adequação a etapa da Análise de Conteúdo.

O artigo está organizado da seguinte forma: a seção 2 apresenta o referencial, na seção 3 são apresentados trabalhos relacionados, na seção 4 a metodologia, desenvolvimento do sistema, experimento e resultados e na seção 5 a conclusão.

2. REFERENCIAL TEÓRICO

2.1. Análise de Conteúdo

Conforme descrito em (Mayring, 2000) a análise de conteúdo é uma técnica das ciências da comunicação desenvolvida para analisar meios de comunicação em massa (jornais, rádio). Evoluiu do campo quantitativo como contagem e atribuição de pesos e relacionamento entre os elementos do texto, para uma análise de conteúdo considerando contexto e estruturas de sentido latentes. Assim é utilizada em estudos na Administração como metodologia rígida para análise qualitativa. Para contextualizar, tem-se a definição:

> 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 teoria, o qual determina os aspectos que devem ser filtrados do material. A forma básica adotada no Bureau é a Estruturação, definida em (Mayring, 2000) com o objetivo de estabelecer um recorte do material na base de critérios pré-estabelecidos. E as etapas para o processo como (Bardin, 2006) são: pré-análise, exploração do material, tratamento dos resultados, inferência e interpretação. O foco do trabalho está nas tarefas de estabelecimento de categorias da etapa de pré-análise e classificação, que serão descritas a seguir no contexto do Bureau.

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 (Bardin apud Valentim, 2005). Neste contexto, os dados precisam de organização formal que capture as características da cafeicultura para gerar informação útil para tomada de decisão.

No caso do Bureau, as notícias extraídas da *web* se referem a diferentes dimensões do mercado com fatos relevantes que impactam a variação de oferta e demanda do 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 com relevância para consumo. Desta forma, a cadeia produtiva do café é uma referência fundamental como teoria para a definição de categorias para organizar a recuperação e classificação de informações extraídas da *web*, ao passo que através dela é possível delimitar dimensões para análise.

Neste sentido a definição de categorias pelo Bureau é apriorística e tem como interesse a cadeia produtiva do café.

2.2. Cadeia Produtiva do Café e Categorias

Cadeia produtiva ou cadeia de suprimento é um termo usado para indicar uma sequência de estágio de materiais e processos para fabricação de produtos e serviços. Na agroindústria, a cadeia produtiva do café é representada pelo Sistema Agroindustrial, conforme (Farina & Zylbersztajn, 1998). A Tabela 1 mostra a cadeia produtiva do café no Brasil dividida em segmentos.

Pela análise dos componentes interativos da cadeia do café é possível identificar classes, entidades e relacionamentos que vão definir a organização dos dados para recuperação, classificação e análise. A cadeia é divida em segmentos que se relacionam por transações. 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 constitui uma cadeia e fatos relevantes específicos divulgados em forma de notícias na *web*.

Tabela 1: Segmentos da Cadeia Produtiva do Café

Categoria	**Dimensões abordadas / Subcategorias**	**Segmentos / Subcategorias**
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

Fonte: Elaborado pelos Autores

A primeira definição extraída do modelo são as dimensões propostas: Fornecedores de Insumos, Máquinas e Equipamentos, Produção Primária, Primeiro Processamento, Segundo Processamento, Vendedores Nacionais, Compradores Internacionais e Varejo Nacional e Internacional. Estas serão consideradas dimensões de mais alto nível e as informações extraídas são relacionadas a uma delas. Os segmentos dentro de cada segmento geral são considerados subcategorias das categorias geradas pelo respectivo segmento, por exemplo, Produtores de Mudas é subcategoria de Fornecedores. As subcategorias constituem as categorias principais, para as quais a análise é feita e descrita no relatório: Indústria, Produção, Cafeterias e Bebidas.

Assim as notícias coletadas pelo Bureau são classificadas segundo o sistema de categorias apresentado na Tabela 1. O corpus composto pelas notícias categorizadas no processo de análise de conteúdo forma a base de treinamento para os métodos de aprendizado de máquina testados, conforme é apresentado nas próximas seções que discorrem sobre as bases tecnológicas do trabalho.

2.3. Classificação Textual

No sentido mais amplo, 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 semi-estruturados. Estas técnicas incluem algoritmos para identificação de entidades, *text clustering*, análise de sentimentos e categorização de texto para organização de documentos – objeto de estudo neste trabalho. O objetivo é classificar documentos de textos de acordo com categorias pré-definidas (Joachims, 1998). (Sebastiani, 2002) descreve a tarefa como uma função: $\phi : DxC \rightarrow \{T, F\}$, onde $D = \{d_1, d_2, ... d_{|D|}\}$ é conjunto que representa o domínio de documentos (corpus) 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 classificação é supervisionada quando há informação externa sobre a organização dos documentos – um conjunto de dados textuais previamente rotulados com as classes. Neste caso a abordagem com aprendizado de máquina é conveniente, pois, a partir de um processo indutivo, automaticamente cria-se um classificador de texto por aprendizado de um conjunto de documentos pré-classificados.

Entre os métodos de aprendizado de máquina amplamente estudados como classificadores estão: Naïve Bayes, Decision Trees, K-nearest neighbor (KNN), *Support Vector Machines*, Neural Networks, Latent Semantic Analysis, Rocchio's Algorithm, Fuzzy Correlation and Genetic Algorithms. Uma descrição destes métodos com vantagens e desvantagens é apresentada em (Bilski, 2011) e (Baharudin, Lee, & Khan, 2010) que também comenta o efeito da qualidade dos dados no desempenho dos métodos. Os métodos estudados neste trabalho estão descritos na próxima seção.

2.4. Classificadores por Aprendizado de Máquina

Dentre os métodos estudados na literatura (Sebastiani, 2002), (Yang & Liu, 1999), (Kang, Yoo, & Han, 2012), e utilizados como classificadores, são testados neste trabalho os pertencentes a três classes: Naïve Bayes, Árvores de Decisão e *Support Vector Machines*.

Métodos baseados em Árvore de Decisão decompõe hierarquicamente o conjunto de dados de treinamento nas classes pré-definidas (J. Ross Quinlan, 1986) de acordo com características presentes nos valores de seus atributos e decide em qual partição é mais provável que um determinado texto pertença.

Os métodos Bayesianos 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, Dumais, Osman, Platt, & Scholkopf, 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.

O foco deste trabalho são três variações de Bayesianos, duas de Árvore de Decisão e uma SVM, selecionados principalmente pela simplicidade e eficácia destes métodos (Ikonomakis, Kotsiantis, & Tampakas, 2005) e disponibilidade de ferramentas.

3. TRABALHOS RELACIONADOS

Pesquisas exploram classificação textual supervisionada com uso de métodos de aprendizado de máquina em diferentes aplicações com classes pré-definidas para cada contexto especifico, entre elas, (Garcia Adeva, Pikatza Atxa, Ubeda Carrillo, & Ansuategi Zengotitabengoa, 2014) apresenta a aplicação de classificação textual no apoio à fase de triagem de artigos em revisão sistemática médica. Ressalta que a determinação de parâmetros de classificação e seleção de algumas partes dos artigos melhoram os resultados.

Para classificar automaticamente orientações médicas em categorias hierárquicas pré-definidas (Moskovitch et al., 2006) destaca que ajustes nos métodos de treinamento variam a precisão observada nos resultados entre 44% a 60%.

Em uma comparação para classificar textos acadêmicos em quatro categorias: Engenharia Civil, Ciências Sociais, Química e Psicologia, (Venegas, 2007) mostra que Multinomial Naïve Bayes tende a extrair mais informações relevantes do texto do que atingir maior precisão na classificação em relação enquanto SVM tende a precisão.

(Torii et al., 2011) utiliza Naïve Bayes e SVM na classificação de artigos relevantes publicados na web para detectar epidemias. O experimento mostra que o desempenho dos métodos é satisfatório para a tarefa e ambos dependem de características do dado de treinamento e teste.Naïve

Classificação textual também é explorada em Análise de Sentimento para identificar a polaridade do texto, neste caso as categorias são: positiva, negativa ou neutra (Das & Chen, 2007), (Aparna, Ingita, Apurva, Karishma, & Suneet Kumar, 2011), (Wilson, Wiebe, &

Hoffmann, 2009), (Aparna, et al., 2011), ou reconhecer classes mais específicas como raiva e aversão e sua intensidade (Reyes, Rosso, & Buscaldi, 2009), (Khoo, Nourbakhsh, & Na, 2012), (Neviarouskaya, Prendinger, & Ishizuka, 2011).

Os trabalhos apontam que o desempenho dos métodos depende de configurações prévias, pré-processamento e características dos dados. Não há uma abordagem definitiva apontada como eficiente para toda aplicação, o que faz da classificação textual uma área de pesquisa abrangente e em constante estudo.

A construção de sistemas computacionais que lidam com informação textual (não estruturada) não é trivial. São vários desafios que aumentam a complexidade conforme (Dey & Haque, 2009), (Pang & Lee, 2008), (Liu, 2010), como subjetividade e ambiguidade de textos em linguagem natural, informalidade e ruídos: erros gramaticais, pontuação imprópria, abreviações, gírias e palavras erradas, além do maciço volume de dados, necessidade de interpretar a informação em tempo e diversidade de línguas.

Além de uma arquitetura computacional, é necessário um modelo conceitual que defina o que deve ser extraído da *web* e como o resultado deve ser classificado para fazer sentido e produzir conhecimento para a aplicação específica.

As definições e desenvolvimento do trabalho no contexto proposto são descritos na próxima seção.

4. METODOLOGIA – *DESIGN SCIENCE RESEARCH*

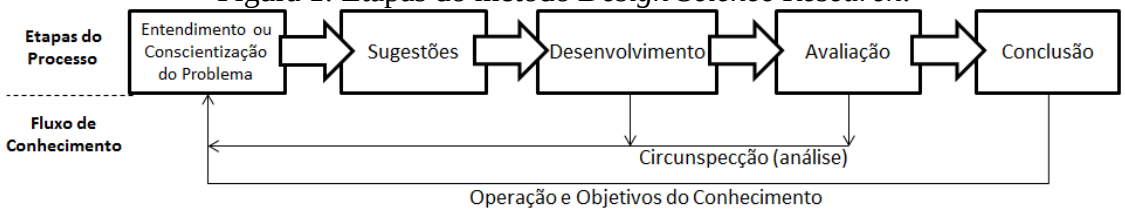
A natureza experimental do trabalho inclui o desenvolvimento de um sistema para coleta e comparação de métodos de classificação, assim é adotado *Design Science Research* como método de pesquisa. A ideia central é que a aquisição de conhecimento e a solução de um problema acontecem pela construção e aplicação de um artefato para um contexto do problema específico. O artefato neste trabalho é o sistema de coleta e classificação automática de notícias para apoio a análise de conteúdo.

Para (Vaishnavi & 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.

O trabalho de (De Sordi, Meireles, & 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.

As etapas deste trabalho seguem o modelo proposto por (Vaishnavi & Kuechler, 2004) aperfeiçoado de (Takeda, Veerkamp, & Yoshikawa, 1990) conforme Figura 1.

Figura 1: Etapas do método *Design Science Research*.



Fonte: Adaptado de (Takeda, et al., 1990) e (Vaishnavi & Kuechler, 2004).

A investigação tem início pelo conhecimento de um problema ou oportunidade de pesquisa na etapa de Entendimento ou Conscientização do Problema. Na etapa de sugestão são elaborados um ou mais modelos de tentativa para a resolução do problema a partir da existência de conhecimento/teoria de base sobre o problema. O artefato é construído na etapa de Desenvolvimento usando técnicas específicas ao artefato criado. Depois de construído é

avaliado em função dos critérios propostos. E na etapa de conclusão são consolidados e registrados os resultados da pesquisa. Cada etapa é brevemente descrita a seguir.

4.1.Entendimento do Problema

O Bureau de Inteligência Competitiva do Café é um projeto do Centro de Inteligência em Mercados da Universidade Federal de Lavras que utiliza dados da *web* para auxiliar a produção de relatórios com delineamento de cenários para tomada de decisões na cafeicultura.

Uma equipe, formada por especialistas, monitora notícias sobre o mercado diariamente na *web* com uma de ferramentas de busca e as classifica de acordo com relevância para áreas da cadeia produtiva do café. Com os especialistas, é possível verificar o que é extraído da *web*, como o resultado deve ser classificado e o que é analisado. Uma dimensão analisada para a Inteligência Competitiva é o ambiente, caracterizado pela cadeia produtiva do café.

A pesquisa é por termos via ferramenta de busca do Google. Os resultados relevantes são rotulados de acordo com as categorias da cadeia produtiva do café pré-definidas em análise de conteúdo e armazenadas em um banco de dados para produção de relatórios.

O banco de dados tem uma tabela de notícias com data, link, texto, fonte e categoria que indica o setor da cadeia produtiva para o qual a notícia é relevante. É composto por 2952 notícias em inglês, coletadas manualmente da web por especialistas do Bureau no período de 01/01/2011 a 31/12/2014. São armazenados título, fonte, conteúdo e categoria. São 414 notícias como categoria Bebidas, 1093 como Cafeterias, 672 como Indústria e 706 como Produção. A categoria é atribuída após a leitura do texto pelo especialista.

Os relatórios são produzidos mensalmente, porém todo o processo é restrito a capacidade dos especialistas para coletar de diversas fontes, ler e categorizar as notícias relevantes, o que consome tempo e recursos humanos. Nem sempre todas as categorias têm notícias suficientes para produzir uma análise abrangente o que pode comprometer a qualidade do relatório e consequentemente seu propósito.

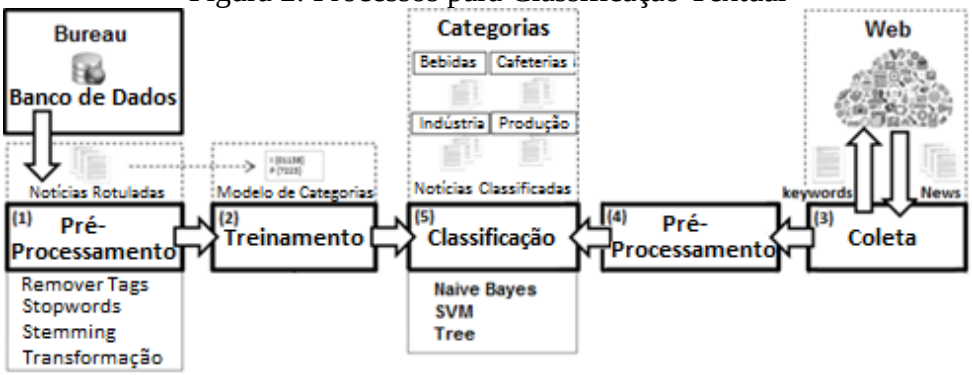
Neste cenário, temos o problema enunciado como pergunta: A partir da definição de categorias é possível promover uma etapa da análise de conteúdo através de um sistema de coleta e classificação automática de notícias da web?

A partir do conhecimento do problema, é apresentado um modelo para tentativa de resolução na etapa de sugestão a seguir.

4.2. Sugestão

A sugestão parte do conhecimento/teoria sobre os processos de mineração de dados com etapas até a classificação textual, ilustrados na Figura 2, e tem como referencia modelos tradicionais para extração de conhecimento (Moraes, Valiati, & Gaviao Neto, 2013), (Bramer, 2013).

Figura 2: Processos para Classificação Textual



Fonte: Elaborada pelos autores

A primeira etapa (1) consiste em recuperar notícias do banco de dados do Bureau para treinamento dos classificadores que irão realizar a categorização de novas notícias extraídas pelo sistema. Esta etapa de pré-processamento inclui limpeza e transformação. A limpeza tem como objetivo preparar o conteúdo apenas com o essencial para as próximas etapas, para isso realiza a retirada de *tags HTML*, *scripts*, ruídos e *stopwords*, reduz a variação dos termos para uma representação por stemming e transforma o conteúdo em uma representação adequada para entrada de cada classificador. O texto é convertido em um vetor de palavras (*bag of words*) como uma representação numérica. Todas as notícias passam por este processo para serem utilizadas na fase de treinamento.

Na etapa de treinamento (2) os vetores gerados pelo texto de cada notícia são associados a sua categoria pré-definida, desta forma cada método de classificação fica pronto para classificar por comparação estatística as próximas notícias que receber.

Para recuperar notícias da web a partir de termos pré-definidos, na etapa (3), o sistema utiliza uma ferramenta de busca. As notícias resultantes são pré-processadas em (4) pelo mesmo procedimento aplicado em (1) e entregues para o classificador.

Na fase de classificação (5) as notícias recuperadas da web são automaticamente classificadas de acordo com as categorias aprendidas na fase de treinamento. Ao final do processo os dados coletados estarão organizados de acordo com as categorias pré-definidas e é possível acrescenta-los a base de treinamento, caso seja interessante para a aplicação.

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 segue as etapas do modelo apresentado na seção 4.2.

4.3.1. Pré-processamento

Na fase de pré-processamento, para limpeza do texto utilizou-se a biblioteca de classes em Java, Apache Tika (Mattmann & Zitting, 2011), que contem funções específicas para este tipo de tratamento de texto. Este procedimento é necessário antes do treinamento, pois como as notícias são copiadas manualmente da web pelos especialistas, 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*, em Inglês, considerada está disponível em http://www.ranks.nl/stopwords. O algoritmo para a tarefa de *stemming* é o Snowball (Porter, 2001), popularmente utilizado para língua inglesa.

No processo de transformação da notícia em um vetor de palavras o filtro adequado na ferramenta WEKA é o *StringToWordVector*. Sua função é converter texto em um conjunto de atributos representando a ocorrência de cada palavra no texto. Este conjunto é armazenado em um vetor 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, & Schütze, 2008): $TF - IDF_{t,d} = TF_{t,d} \times IDF_t$, onde: $IDF_t = \log\frac{N}{DF_t}$; $TF_{t,d}$ é número de ocorrências do termo t no documento d; N é número de documentos no conjunto; DF_t é número de documentos no conjunto que contem o termo t.

A etapa de pré-processamento entrega o texto em representação adequada para

treinamento e classificação, e melhora o desempenho dos classificadores (Haddi, Liu, & Shi, 2013).

4.3.2. Treinamento e Classificação

Os vetores são gerados a partir das notícias na fase de transformação e as categorias definidas pela análise de conteúdo. A classificação supervisionada é dividida em duas fases distintas: treinamento e classificação. Na fase de treinamento, um algoritmo com a técnica específica é treinado com um conjunto de documentos previamente classificados. Na fase de classificação o algoritmo classifica novos documentos de acordo com as categorias aprendidas na fase anterior.

Para as fases de treinamento e classificação, o banco de dados é dividido em dois conjuntos de dados: $\Omega_t = \{d_1, .. d_{|\Omega_t|}\}$, onde $|\Omega_t|$ representa o número de notícias compreendido entre duas datas para treinamento. E o conjunto: $\Omega_c = \{d_1, .. d_{|\Omega_c|}\}$, onde, $|\Omega_c|$ representa o número de notícias compreendido entre duas datas para classificação, usado para avaliar o desempenho dos classificadores.

As categorias para classificação são as quatro definidas no Quadro 1 representadas pelo conjunto: *C* = {*"Indústria", "Produção", "Cafeterias", "Bebidas"*} e os algoritmos classificadores são representados pelo conjunto: $\phi = \{\phi_1, ..., \phi_7\}$.

Foram utilizadas as versões dos classificadores codificados na ferramenta WEKA com parâmetros padrão, a saber: Bayesianos – Naïve Bayes (John & Langley, 1995), Naïve Bayes Multinomial (McCallum & Nigam, 1998), Complement Naïve Bayes (Rennie, Shih, Teevan, & Karger, 2003), Discriminative Multinomial Naïve Bayes (J. Su, Zhang, Ling, & Matwin, 2008); Árvore de Decisão – J48 (John Ross Quinlan, 1993), Random Tree e Support Vector Machines – SMO (Platt, 1998).

O experimento foi executado em uma aplicação programada em Java, desenvolvida para conectar ao banco de dados do Bureau, pesquisar os conjuntos de notícias, treinar algoritmos da WEKA pelo uso de sua interface de programação e utilizá-los para classificar conjuntos de notícias.

Com a descrição dos parâmetros anteriores, o procedimento é: 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, foram considerados três conjuntos para treinamento e teste no experimento, descritos na fase de avaliação. Os conjuntos de notícias classificados por cada método foram comparados com a classificação realizada pelos especialistas e verificado o índice de acerto geral por medidas comumente utilizadas em classificação textual que serão descritas em 4.4.

4.3.3. Coleta

Os especialistas do Bureau utilizam o Google News para pesquisar diariamente as notícias na web, utilizando uma lista de termos, definida pelos profissionais, que inclui nomes dos principais agentes da cadeia do café como empresas – Starbucks, Nestle, Green Mountain, Lavazza, entre outras, países concorrentes, e termos comuns no domínio de cada categoria da cadeia produtiva como: *coffee, drink, coffee shop, production,* entre outros.

Assim, para automatizar a tarefa de coleta de dados, o sistema percorre um arquivo com a lista de termos pré-definidos pelos especialistas e, através da interface do Google News, recupera notícias para cada termo da lista. Após etapa de limpeza no pré-processamento, cada notícia é classificada em uma categoria utilizando o classificador já treinado com as notícias existentes no banco de dados.

4.4. Avaliação

Conforme (Sebastiani, 2002) o problema clássico de classificação textual tem característica subjetiva e não normalizável e a avaliação experimental usualmente mede eficácia. Assim, foram adotadas as medidas de precisão: $\pi_i = \frac{TPi}{TPi+FPi}$ e $\rho_i = \frac{TPi}{TPi+FNi}$, onde TP_i = Verdadeiro Positivo, mesma categoria C_i pelo classificador e especialista; FP_i = Falso Positivo, classificador atribuiu a categoria C_i e especialista não; FN_i = Falso Negativo, classificador não atribuiu a categoria C_i e especialista sim. Todos para uma categoria C_i. E para obter estimativas de π e ρ localmente, para cada categoria: $\pi^{\mu} = \frac{\sum_{i=1}^{|C|} TPi}{\sum_{i=1}^{|C|}(TPi+FPi)}$ e $\rho^{\mu} = \frac{\sum_{i=1}^{|C|} TPi}{\sum_{i=1}^{|C|}(TPi+FNi)}$, onde μ indica a estimativa média. E globalmente pela fórmula: $\pi^{M} = \frac{\sum_{i=1}^{|C|} \pi_i}{|C|}$ e $\rho^{M} = \frac{\sum_{i=1}^{|C|} \rho_i}{|C|}$, onde $|C|$ é o número de categorias.

Desta forma é possível observar o comportamento dos classificadores para categorias com diferentes quantidades de notícias para treinamento.

Para a avaliação, foram realizados experimentos em três etapas: uma para verificar o desempenho dos classificadores com conjunto de treinamento constante em um conjunto de teste de um ano, outra com conjunto de treinamento variável e teste de um mês e outra para coleta de um mês de notícias.

4.4.1. Experimento com Base de Treinamento Constante

O primeiro teste teve como objetivo analisar o desempenho dos classificadores em um conjunto de teste maior, formado por um ano de notícias, a partir do treinamento com uma base constante de notícias. Nesta etapa, o conjunto de treinamento Ω_t foi composto por 2507 notícias em inglês coletadas entre 01/01/2011 e 31/12/2013. E o conjunto de teste Ω_c formado por 387 notícias coletadas no período de 01/01/2014 a 31/12/2014, também em inglês, distribuídas em categorias conforme Tabela 1 da Seção 2.2.

As notícias do ano de 2014 foram classificadas utilizando cada método conforme pseudo-código da seção 4.3.2 e comparadas com a classificação do especialista em métricas apresentadas nesta seção. O resultado distribuído por categorias é apresentado na Tabela 2.

Tabela 2: Resultado da Classificação distribuído por categorias

	Categorias																				**Total**			
	Bebidas (Ω_t : 402; Ω_c 12)					**Cafeterias** (Ω_t : 909; Ω_c : 184)					**Indústria** (Ω_t : 577; Ω_c : 95)					**Produção (619 : 87)** (Ω_t : 619; Ω_c : 87)					**Microavg**		**Macroavg**	
Class.	**TP**	**FP**	**FN**	**Π**	**P**	**TP**	**FP**	**FN**	**π**	**P**	**TP**	**FP**	**FN**	**Π**	**P**	**TP**	**FP**	**FN**	**π**	**ρ**	**π**	**ρ**	**π**	**ρ**
R.Tree	3	**46**	**9**	0,06	0,25	112	46	**72**	0,71	0,61	43	**42**	52	0,51	0,45	60	**26**	**27**	0,70	0,69	0,58	0,58	0,49	0,50
J48	8	14	4	0,36	0,67	145	**67**	39	0,68	0,79	32	29	**63**	0,52	0,34	69	14	18	0,83	0,79	0,67	0,67	0,60	0,65
DMNBtext	**11**	16	1	0,41	**0,92**	**173**	31	11	0,85	**0,94**	59	4	36	**0,94**	0,62	82	2	5	**0,98**	0,94	0,86	0,86	0,79	0,86
NaïveBayes	**11**	10	1	0,52	**0,92**	158	19	26	**0,89**	0,86	**76**	13	19	0,85	**0,80**	83	8	4	0,91	0,95	**0,87**	**0,87**	0,80	**0,88**
NBMult.	**11**	8	1	**0,58**	**0,92**	162	27	22	0,86	0,88	68	9	27	0,88	0,72	85	8	2	0,91	0,98	0,86	0,86	**0,81**	0,87
Comp.NB	**11**	10	1	0,52	**0,92**	164	23	20	0,88	0,89	65	6	30	0,92	0,68	**87**	12	0	0,88	**1,00**	0,87	0,87	0,80	0,87
SMO	**11**	16	1	0,41	**0,92**	158	46	26	0,77	0,86	59	16	36	0,79	0,62	70	2	17	0,97	0,80	0,79	0,79	0,74	0,80

Fonte: Elaborado pelos autores

Os classificadores Bayesianos obtiveram melhores resultados por categoria e consequentemente quanto ao número total de acertos – classificações iguais as dos especialistas, conforme ilustra o gráfico na Figura 3. E também melhor π local.

O treinamento com a base de dados de 2011 a 2013 resultou em uma classificação com índice de acerto acima de 80% para as quatro variações de Naïve Bayes, conforme gráfico mostrado na Figura 7, com destaque para o método Naïve Bayes com índice de acerto de 86,77%. O método SMO da classe Support Vector Machines teve desempenho satisfatório, mas inferior aos da classe Naïve Bayes.

Há uma diferença de desempenho em relação a categorias específicas. A Tabela 2 mostra que o conjunto de teste Ω_c possui apenas 12 notícias classificadas como Bebidas, neste caso o método Naive Bayes Mutinomial apresentou melhor desempenho com menor número de FP. E apesar de DMNBtext e SMO terem o mesmo número de TP apresentaram alto FP o que reduz a precisão local π para a categoria.

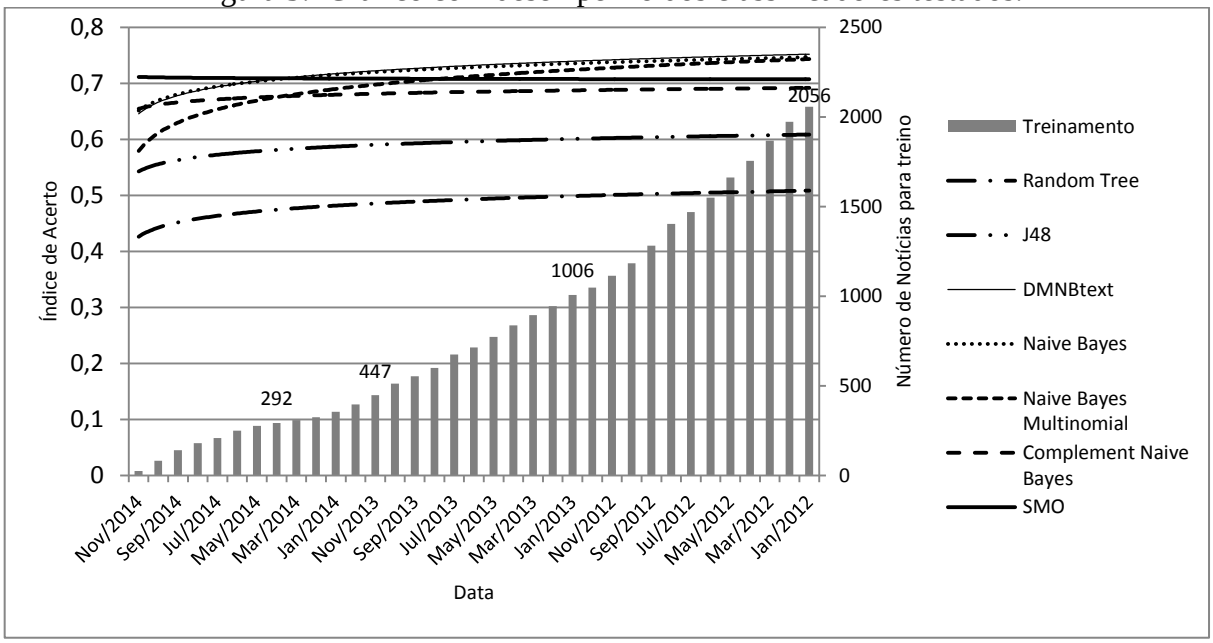
Os métodos baseados em Árvore apresentam desempenho inferior no resultado geral, e em categorias específicas, principalmente pelo elevado número e FP para as categorias.

4.4.2. Experimento com Base de Treinamento Variável

O objetivo é verificar o comportamento dos clasificadores a medida que a base de treinamento aumenta. Assim, o conjunto de teste Ω_c foi composto pelas notícias extraídas no período de um mês – período usado para produzir um relatório – 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 2056 notícias coletadas de 01/01/2012 até 30/11/2014 para treinamento.

Desta forma percebe-se o comportamento dos classificadores à medida que a massa de dados de treinamento aumenta, conforme Figura 3.

Figura 3: Gráfico com desempenho dos classificadores testados.



Fonte: Elaborado pelos autores

Para facilitar a visualização e comparação dos classificadores, as curvas com os resultados de cada classificador apresentadas na Figura 4 foram suavizadas por logaritmo. À medida que os dados de treinamento aumentam o desempenho dos classificadores também aumenta, entretanto após um nível aproximado de 400 a 500 notícias de treinamento o aumento passa a ser menos significativo. Os métodos baseados em Árvore ficam no máximo em 60%. Já os métodos Bayesianos continuam aumentando mais discretamente o desempenho à medida que os dados de treinamento aumentam, mas se estabilizam acima de 75% e abaixo de 80%. O método Support Vector Machines tem a peculiaridade de apresentar um desempenho estável em 70% mesmo para poucas amostras de treinamento.

Com o aumento do conjunto de treinamento os índices de acerto dos algoritmos Naïve Bayes subiram e se estabilizaram em níveis significativos 73,9%. Cabe ressaltar que o teste foi realizado para um mês específico e como as notícias de treinamento e teste podem ser desbalanceadas quanto às categorias e carregam a visão e julgamento dos especialistas, variam quanto as suas características o que sugere maior investigação para aumento de precisão o que não é escopo deste trabalho.

4.4.3. Experimento de Coleta e Classificação

Nesta etapa o objetivo foi classificar notícias coletadas pelo sistema e comparar com as coletadas pelos especialistas no mesmo período. Para isso foi incluído um módulo no sistema para buscar automaticamente notícias utilizando a interface do Google News pelo link: http://ajax.googleapis.com/ajax/services/search/news?v=1. Com este recurso, o sistema consulta as mesmas palavras e fontes pesquisadas manualmente pelos especialistas, classifica as notícias recuperadas e armazena no banco de dados. Para este teste, foi selecionado apenas o classificador DMNBtext dado seu desempenho satisfatório nos testes anteriores.

Como base de treinamento foram utilizadas 2901 notícias de 01/01/2011 até 31/01/2015, coletadas pelos especialistas. A base de teste foi resultante da consulta automática realizada no dia 17/06/2015 com termos e fontes pesquisadas manualmente em Abril pelos especialistas para produção do relatório de Maio.

Foi realizada uma consulta cruzada de cada termo com cada palavra que descreve a fonte. E para avaliação foram consideradas notícias recuperadas com data de publicação entre 01/04/2015 a 30/04/2015.

Os especialistas coletaram 18 notícias em Abril, 13 para categoria Cafeterias e 5 para Indústria, todas em Inglês, enquanto o sistema recuperou automaticamente da web 176 notícias em Inglês com data de publicação em Abril.

Como não há uma base negativa para treinamento, todas foram classificadas em alguma categoria conhecida, gerando ruído na amostra que compõe o conjunto de teste. Das 175 notícias coletadas pelo sistema 108 foram consideradas irrelevantes para o contexto e 66 notícias foram consideradas relevantes pelos especialistas. Destas, 53 não foram coletadas da web no processo manual para análise e 13 também foram coletadas pelos especialistas. Porém há uma mesma notícia publicada em duas fontes diferentes, portanto tem-se 12 notícias em comum na coleta entre sistema e especialistas.

O problema das notícias indesejadas sugere a criação de uma base de notícias irrelevantes para treinamento, o que pode ser realizado pelos especialistas após coleta e classificação, tornando o processo semi-supervisionado até que se tenha uma base adequada para que o classificador elimine notícias que não pertencem ao contexto.

Outra questão a ser observada é que 6 notícias coletadas manualmente pelos especialistas não foram recuperadas pelo sistema. Assim, é importante que o processo seja diário evitando que notícias relevantes sejam perdidas em consultas com datas posteriores a sua publicação.

Entretanto, como o objetivo neste trabalho é comparar o resultado do classificador com a classificação do especialista a verificação foi realizada com notícias recuperadas por ambos. A tabela XX mostra a comparação da classificação realizada pelo sistema para as 12 notícias e a classificação realizada pelos especialistas para as mesmas notícias. O método DMNBtext apresentou resultado com 10 acertos e 2 erros.

Tabela 2: Resultado da Classificação distribuído por categorias

Link	Data de Publicação	DMNBtext	Especialista
http://www.kpbs.org/news/2015/apr/17/espresso-in-orbit-spacex-craft-brings-coffee/	17/04/2015	Cafeterias	Indústria
http://www.vendingmarketwatch.com/news/12064683/melitta-usa-launches-keurigr-20r-compatible-single-serve-coffee	14/04/2015	Indústria	Indústria
http://www.balkans.com/open-news.php?uniquenumber=202517	02/04/2015	Cafeterias	Indústria
http://www.bakeryinfo.co.uk/news/fullstory.php/aid/14105/Coffee_chain_proposes__nap_puccino.html	01/04/2015	Cafeterias	Cafeterias
http://www.vancitybuzz.com/2015/04/tim-hortons-single-origin-coffee/	01/04/2015	Cafeterias	Cafeterias
http://www.chicagobusiness.com/article/20150403/BLOGS09/150409907/dollop-coffee-is-planning-new-digs-for-this-summer	03/04/2015	Cafeterias	Cafeterias
http://www.thenational.ae/business/the-life/independent-roasters-and-cafes-blossom-in-uae	06/04/2015	Cafeterias	Cafeterias
https://ca.shine.yahoo.com/blogs/shine-on/you-told-us--tim-hortons-should-go-back-to-the-basics-163538503.html	06/04/2015	Cafeterias	Cafeterias
http://www.businessinsider.com/caffeine-consumption-by-age-2015-4	06/04/2015	Cafeterias	Cafeterias
http://www.businessinsider.in/Heres-how-much-caffeine-people-consume-at-every-age/articleshow/46829694.cms	06/04/2015	Cafeterias	Cafeterias
http://houston.eater.com/2015/4/6/8354581/downtown-houston-hilton-americas-new-starbucks-concept	06/04/2015	Cafeterias	Cafeterias
http://www.vendingmarketwatch.com/news/12069943/marley-coffee-expands-distribution-in-chile	30/04/2015	Cafeterias	Cafeterias

Fonte: Elaborado pelos autores.

O desempenho do método DMNBtext representa 83% de acerto das notícias coletadas, o que é satisfatório para o propósito do sistema. Considerando a avaliação de todo o sistema, incluindo o processo de coleta, o resultado representa 55% de desempenho, 10 acertos em 18 notícias coletadas pelos especialistas, o que aponta para necessidade de melhoria na etapa de coleta de notícias.

5. CONCLUSÃO

Os resultados mostram um desempenho superior dos classificadores Bayesianos para os dados do Bureau. Na classificação das notícias do ano de 2014 eles apresentaram resultado superior a 85% de categorização igual a dos especialistas com um conjunto de treinamento formado por 2507 notícias. E também no teste de comportamento quanto à variação do número de amostras, apresentam uma curva ascendente no número de acertos.

Entretanto, não é um resultado definitivo ao passo que o comportamento dos classificadores se altera com características do conjunto de treinamento, pré-processamento bem como parâmetros dos próprios classificadores. Porém, é satisfatório para o objetivo do trabalho, pois comprova, através de um artefato, que é viável a utilização dos algoritmos de aprendizado de máquina para automatizar o processo de categorização de notícias e contribuir para análise de conteúdo realizada pelos especialistas do Bureau, permitindo resolução de um problema relevante para o contexto do negócio.

O trabalho mostra a utilidade e eficácia do artefato através de um método rigoroso para construção e validação respeitando as regras de análise do Bureau. Além disso, a arquitetura da aplicação em desenvolvimento é flexível para que novos classificadores possam ser testados para aprimorar o sistema e permite que novas categorias sejam criadas de acordo

com a análise dos especialistas, o que representa ganho de escala para análise de grandes volumes de dados.

O trabalho corrobora a viabilidade, aperfeiçoamento e continuidade do sistema em desenvolvimento e indica que a precisão pode ser aumentada adotando-se um processo semi-supervisionado onde ajustes são realizados pelos especialistas após categorização automática do conjunto de notícias extraídas da *web*, o que representa um avanço pela redução de recursos necessários para todo o processo. Os resultados sugerem novos testes combinando os métodos com o objetivo de aumentar a precisão, uma vez que o comportamento altera quando analisados por categoria.

Com o objetivo de eliminar ruídos no processo de coleta é necessário criar uma base de notícias irrelevantes para treinamento do classificador, o que pode ser realizado pelos especialistas enquanto usam o sistema. Outro fator é a realização de pesquisa diária para cobrir notícias recentes e evitar que sejam desprezadas, o que também contribui para aumentar a chance de coletar as mesmas notícias selecionadas pelos especialistas.

Trabalhos futuros incluem as tarefas citadas para coleta, a inclusão de mecanismos para extrair entidades, categorias mais específicas como fusões entre empresas, eventos climáticos e explorar a análise de sentimento correlacionada com variação de preço do café no mercado, visando um suporte mais robusto não só para gerenciamento de dados para produção de relatórios, mas também geração de conhecimento para criação de inteligência competitiva para a cafeicultura brasileira.

REFERÊNCIAS

Aparna, T., Ingita, S., Apurva, S., Karishma, S., & Suneet Kumar, G. (2011). Literature Survey on Design and Implementation of Processing Model for Polarity Identification on Textual Data of English Language. *International Journal of Computer Science Issues, 8*(6), 309.

Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology, 1*(1), 4-20.

Bardin, L. (2006). Análise de conteúdo (L. de A. Rego & A. Pinheiro, Trads ed.): Lisboa: Edições 70. (Obra original publicada em 1977).

Bilski, A. (2011). A review of artificial intelligence algorithms in document classification. *International Journal of Electronics and Telecommunications, 57*(3), 263-270.

Bramer, M. (2013). *Principles of data mining*: Springer.

Collier, N., Doan, S., Kawazoe, A., Goodwin, R. M., Conway, M., Tateno, Y., et al. (2008). BioCaster: Detecting public health rumors with a Web-based text mining system. *Bioinformatics, 24*(24), 2940-2941.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273-297.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science, 53*(9), 1375-1388.

De Sordi, J. O., Meireles, M., & Sanches, C. (2011). DESIGN SCIENCE APLICADA ÀS PESQUISAS EM ADMINISTRAÇÃO: REFLEXÕES A PARTIR DO RECENTE HISTÓRICO DE PUBLICAÇÕES INTERNACIONAIS DOI: 10.5773/rai. v8i1. 770. *RAI: revista de administração e inovação, 8*(1), 10-36.

Dey, L., & Haque, S. (2009). Opinion mining from noisy text data. *International Journal on Document Analysis and Recognition (IJDAR), 12*(3), 205-226.

Farina, E. M. M. Q., & Zylbersztajn, D. (1998). *Competitividade no agribusiness brasileiro*. Paper presented at the PENSA.

Feuerriegel, S., & Neumann, D. (2013). News or Noise? How News Drives Commodity Prices.

Garcia Adeva, J. J., Pikatza Atxa, J. M., Ubeda Carrillo, M., & Ansuategi Zengotitabengoa, E. (2014). Automatic text classification to support systematic reviews in medicine. *Expert Systems with Applications, 41*(4), 1498-1508.

Haddi, E., Liu, X., & Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia Computer Science, 17*, 26-32.

Hearst, M. A., Dumais, S., Osman, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *Intelligent Systems and their Applications, IEEE, 13*(4), 18-28.

Holmes, G., Donkin, A., & Witten, I. H. (1994). *Weka: A machine learning workbench.* Paper presented at the Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on.

Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS Transactions on Computers, 4*(8), 966-974.

Joachims, T. (1998). *Text categorization with support vector machines: Learning with many relevant features*: Springer.

John, G. H., & Langley, P. (1995). *Estimating continuous distributions in Bayesian classifiers.* Paper presented at the Proceedings of the Eleventh conference on Uncertainty in artificial intelligence.

Junqué De Fortuny, E., De Smedt, T., Martens, D., & Daelemans, W. (2012). Media coverage in times of political crisis: A text mining approach. *Expert Systems with Applications, 39*(14), 11616-11622.

Kang, H., Yoo, S. J., & Han, D. (2012). Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications, 39*(5), 6000-6010.

Khoo, C. S.-G., Nourbakhsh, A., & Na, J.-C. (2012). Sentiment analysis of online news text: a case study of appraisal theory. *Online Information Review, 36*(6), 858-878.

Lee, C. J., Wu, Y. C., & Chen, Y. C. (2012). Building news sentiment indicators for stock marketing application. *International Journal of Advancements in Computing Technology, 4*(2), 103-110.

Li, N., Liang, X., Li, X., Wang, C., & Wu, D. D. (2009). Network environment and financial risk using machine learning and sentiment analysis. *Human and Ecological Risk Assessment, 15*(2), 227-252.

Liu, B. (2010). Sentiment analysis: A multi-faceted problem. *IEEE Intelligent Systems, 25*(3), 76-80.

Malouf, R., & Mullen, T. (2008). Taking sides: User classification for informal online political discourse. *Internet Research, 18*(2), 177-190.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1): Cambridge university press Cambridge.

Mattmann, C., & Zitting, J. (2011). *Tika in Action*: Manning Publications Co.

Mayring, P. (2000). Qualitative content analysis. (7 ed., Vol. 2). Forum Qual Soc Res.

McCallum, A., & Nigam, K. (1998). *A comparison of event models for naive bayes text classification.* Paper presented at the AAAI-98 workshop on learning for text categorization.

McCallum, A., Nigam, K., Rennie, J., & Seymore, K. (1999). *A machine learning approach to building domain-specific search engines.* Paper presented at the IJCAI.

Moraes, R., Valiati, J. F., & Gaviao Neto, W. P. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications, 40*(2), 621-633.

Moskovitch, R., Cohen-Kashi, S., Dror, U., Levy, I., Maimon, A., & Shahar, Y. (2006). Multiple hierarchical classification of free-text clinical guidelines. *Artificial Intelligence in Medicine, 37*(3), 177-190.

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2011). SentiFul: A lexicon for sentiment analysis. *IEEE Transactions on Affective Computing, 2*(1), 22-36.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1-2), 1-135.

Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines.

Porter, M. (2001). Snowball: A language for stemming algorithms.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning, 1*(1), 81-106.

Quinlan, J. R. (1993). *C4. 5: programs for machine learning* (Vol. 1): Morgan kaufmann.

Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003). *Tackling the poor assumptions of naive bayes text classifiers.* Paper presented at the ICML.

Reyes, A., Rosso, P., & Buscaldi, D. (2009). Humor in the blogosphere: First clues for a verbal humor taxonomy. *Journal of Intelligent Systems, 18*(4), 311-331.

Schumaker, R. P., Zhang, Y., Huang, C.-N., & Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems, 53*(3), 458-464.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys, 34*(1), 1-47.

Su, J., Zhang, H., Ling, C. X., & Matwin, S. (2008). *Discriminative parameter learning for Bayesian networks.* Paper presented at the Proceedings of the 25th international conference on Machine learning.

Su, Q., Zheng, Y., & Swen, B. (2008). Combined approach of web mining and semantic annotation for identifying product features in customer reviews. *Journal of Computational Information Systems, 4*(3), 1047-1054.

Takeda, H., Veerkamp, P., & Yoshikawa, H. (1990). Modeling design process. *AI magazine, 11*(4), 37.

Torii, M., Yin, L., Nguyen, T., Mazumdar, C. T., Liu, H., Hartley, D. M., et al. (2011). An exploratory study of a text classification framework for Internet-based surveillance of emerging epidemics. *International Journal of Medical Informatics, 80*(1), 56-66.

Vaishnavi, V., & Kuechler, W. (2004). Design research in information systems.

Valentim, M. L. P. (2005). *Métodos qualitativos de pesquisa em Ciência da Informação*: Polis.

Venegas, R. (2007). Academic text classification based on lexical-semantic content. *Revista Signos, 40*(63), 239-271.

Weiss, S. M., Indurkhya, N., Zhang, T., & Damerau, F. (2010). *Text mining: predictive methods for analyzing unstructured information*: Springer.

Wilson, T., Wiebe, J., & Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics, 35*(3), 399-433.

Xu, K., Liao, S. S., Li, J., & Song, Y. (2011). Mining comparative opinions from customer reviews for Competitive Intelligence. *Decision Support Systems, 50*(4), 743-754.

Yang, Y., & Liu, X. (1999). *A re-examination of text categorization methods.* Paper presented at the Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.

Yu, X., Liu, Y., Huang, J. X., & An, A. (2012). Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain. *Ieee Transactions on Knowledge and Data Engineering, 24*(4), 720-734.