INSTITUCIONAL/IFSP	PROJETO DE PESQUISA

TÍTULO DO PROJETO: *Text Mining* na classificação de notícias – Parte 2											
Área do Conhecimento (Tabela do CNPq):	**1**	**.**	**0**	**3**	**.**	**0**	**3**	**.**	**0**	**0**	**-**	**6**

1. RESUMO

Esta é a continuação de um projeto de Iniciação Científica PIBIFSP em desenvolvimento pelo orientador e por aluno do curso de Análise e Desenvolvimento de Sistemas. O trabalho está no escopo de *Text Mining*, definido pela extração de informações relevantes de textos. O processo de *Text Mining* possui várias etapas, partindo da obtenção dos documentos, tratamento dos textos, até a criação de modelos e aplicações. No primeiro ano, o foco foi a obtenção e o tratamento de quase 1,5 milhões de notícias de 1999 até 2017 nas categorias Nacional, Internacional, Política, Economia, Esportes, entre outras, dos sites Diário do Grande ABC e Jornal do Brasil, além da análise da positividade/negatividade das notícias em determinadas categorias. No segundo ano, pretende-se aplicar técnicas de categorização e agrupamento das notícias já extraídas e disponíveis. O objetivo disto é a utilização de procedimentos de modelagem de textos para fazer classificações e agrupamentos para aplicações, das quais destacam-se, *i)* projeção do preço das ações de empresas a partir da positividade/negatividade de notícias e *ii)* identificar as informações e categorizá-las de acordo ao seu potencial favorável/desfavorável de influência para uma organização ou setor empresarial. Esta última aplicação será feita em parceria com pesquisador da PUC-Campinas.

2. FUNDAMENTAÇÃO TEÓRICA

O processo de extrair informações úteis de dados derivados de uma massa de informações gerada a cada momento em algum lugar é chamado de *Data Minining,* ou Mineração de Dados (AHLEMEYER-STUBBE e COLEMAN, 2014). Suas aplicações são as mais diversas, como predizer quando um paciente hospitalizado devido a um ataque do coração terá um segundo evento baseado em informações demográficas, dieta e medidas clínicas; prever o preço de ações em seis meses no futuro, com base em medidas de desempenho de empresas e dados da economia; identificar os números em um código postal escrito à mão em uma imagem digitalizada; estimar a quantidade de glicose no sangue de uma pessoa com diabetes pelo espectro de absorção do sangue da pessoa; identificar os fatores de risco do câncer de próstata, baseado em variáveis clínicas e demográficas (HASTIE, TIBSHIRANI e FRIEDMAN ,2009).

As aplicações, de uma forma geral, são categorizadas em três usos, sendo eles, agrupamento, classificação e regressão. A primeira consiste em agrupar um conjunto de dados em grupos similares de informações de acordo com um objetivo. No geral, não exige uma variável resposta, uma vez que este agrupamento

segue uma função objetivo que pode ser minimizar a variância entre grupos, a distância entre eles ou outra métrica.

O segundo uso consiste em classificar conjuntos de dados em categorias, como níveis de risco, tipos de atividades, categorias de trabalho, entre outros. A classificação exige uma variável resposta para que o método "aprenda" a regra de classificação através de um conjunto de treinamento.

A terceira exige também uma variável resposta para que possa ser feita uma predição de alguma informação com base em um conjunto de variáveis e um histórico de dados para que seja "aprendida" a regra. Por exemplo, a probabilidade de um paciente ter um ataque do coração, o nível de risco de um cliente e a projeção do preço de uma moeda estrangeira.

Para os usos que exigem uma variável resposta, este processo chama-se de aprendizado supervisionado. No processo de modelagem, divide-se os dados em dois conjuntos, sendo um deles o de treinamento e outro de validação. Através do conjunto de treinamento, a regra é aprendida, e através do segundo, verifica-se se a regra continua prevendo adequadamente os dados. Uma situação que deve ser evitada é o *overtitting*, ou super aprendizado da regra, o que se caracteriza por taxas de acerto altíssimas no conjunto de treinamento e taxas baixas no conjunto de validação, o que significa que a regra funciona apenas para o conjunto de treinamento e não tem poder preditivo. Rogers e Girolami (2011) apresentam uma forma de evitar o *overtitting* através do procedimento *K-fold*, que consiste em dividir o conjunto de dados em K partes. Este processo é recorrente e na primeira iteração considera que a 1ª parte é o conjunto de validação e todas as demais (2ª até K-ésima) o conjunto de treinamento. Isto é feito até que a K-ésima parte seja de treinamento e as demais (1ª a (K-1)-ésima) sejam as de treinamento.

Os dados podem ser estruturados, ou seja, estarem disponíveis em uma base de dados com linhas e colunas, sendo cada linha uma observação e cada coluna, uma variável; ou podem ser não estruturados como, por exemplo, um arquivo de texto. No caso de informações não estruturadas de texto, o processo de extração é um *Data Mining* mas também é chamado de *Text Mining* (KAO e POTIT, 2007).

O processo de *Text Mining* contempla diversas etapas, partindo da obtenção dos documentos, tratamento dos textos e o uso de técnicas para categorizar, agrupar ou classificar.

A etapa de tratamento inclui os passos:

1 – Filtering: remoção de caracteres não alfanuméricos, como caracteres de pontuação;

2 – Stemming: redução das palavras para uma forma básica, utilizando como base seu radical;

3 – Conversão das letras para minúsculo;

4 – Stopword Removal: remoção das palavras que são muito utilizadas na língua, porém que sozinhas não trazem significado direto ao texto (como artigos, preposições, entre outros);

5 – Remoção de acentos e sinais diacríticos das palavras.

Segundo Weiss, Indurkhya e Zhang (2015), uma vez que tratamentos tenham sido realizados para tornar os dados numéricos ou categóricos, métodos padrão de data mining podem ser aplicados.

Este processo todo é realizado com o intuito de aplicar situações, como a identificação de e-mails que devem ser classificados como *spam*, um problema de classificação, ou seja, classifica os e-mails em *spam* ou não através de uma análise dos termos mais comuns para cada uma das duas categorias (HASTIE, 2009).

Entre outras aplicações do *Text Mining,* podem ser citadas:

- Identificar buscas mais frequentes na internet para identificar tendências
- Analisar o histórico de postagens em redes sociais
- Identificar se um determinado texto é adequado para menores de idade
- Categorizar a opinião de leitores

Este projeto em específico é a continuação de um projeto de Iniciação Científica PIBIFSP em desenvolvimento pelo professor orientador e por aluno bolsista cursando Análise e Desenvolvimento de Sistemas. No ano 2017, o foco do trabalho foi a extração e o tratamento de textos de notícias de dois sites, o Diário do Grande ABC (www.dgabc.com.br) e Jornal do Brasil (www.jb.com.br). Foram extraídas quase 1,5 milhões de notícias datadas nos anos de 1999 a 2017 em diversas categorias (Nacional, Internacional, Política, Economia, Esportes e outras). A extração dos dados foi feita usando o software *Scrapy* (*Python*) e os dados armazenados em *SQL.* Os tratamentos foram realizados usando o *software R*.

As nuvens de palavras da Figura 1 mostram os termos mais frequentes das notícias da categoria Internacional para o ano de 2016 antes (esquerda) e depois (direita) dos tratamentos. Pode-se ver que antes dos tratamentos, as palavras mais frequentes são “que”, “para”, “uma” e “com”, que não trazem significado sozinhas por se tratarem de conectivos. Já a nuvem de palavras da direita representa melhor as notícias daquele ano, como as eleições dos Estados Unidos, o terremoto na Itália, o ataque terrorista na França, entre outras.

FIGURA 1. Nuvens de palavras antes e após os tratamentos para a categoria Internacional em 2016.

Após as etapas de extração e tratamento dos dados, foram feitas análises de sentimento (positividade e negatividade) das notícias sobre determinados temas.

Como sequência, pretende-se aplicar técnicas de categorização e agrupamento das notícias já extraídas e disponíveis para a modelagem de problemas usando técnicas de regressão, Redes Neurais e K-means, entre outras metodologias. Dentre as principais aplicações alvo, cita-se a projeção do preço das ações de empresas a partir da positividade/negatividade de notícias; a análise de tendência de jornais a respeito sobre determinados temas; a classificação de textos em categorias; e a identificação de informações e categorização de acordo ao seu potencial favorável/desfavorável de influência para uma organização ou setor empresarial.

Esta última aplicação será feita em parceria com pesquisador da Pontifícia Universidade Católica (PUC)- Campinas. Isto porque, no ambiente empresarial também é possível o uso do *Text Mining*. No processo de formulação e desenvolvimento de estratégias empresariais, os gestores estrategistas precisam, dentre outras ações, coletar, analisar e interpretar informações do ambiente externo no qual a organização opera, com a finalidade de identificar, através de oportunidades e ameaças, a forma como esse ambiente influencia e poderia influenciar a estratégia a ser escolhida e, com isso, o desempenho da organização (IRELAND; HOSKISSON; HITT, 2014).

No ambiente externo das organizações, encontram-se os ambientes geral e competitivo. O ambiente geral, ou macroambiente, apresentará informações, embora não gerenciáveis pela organização, mas fortemente influenciadoras das decisões estratégicas. Aspectos sociais-demográficos, econômicos, políticos, legais, culturais, tecnológicos e ecológicos conterão as informações necessárias de serem identificadas, coletadas, monitoradas e avaliadas (BARNEY; HESTERLY, 2011).

Usualmente a busca e monitoramento das informações contidas no ambiente geral é feita através da internet, via sites de jornais, empresas de pesquisa e organizações oficiais dos governos. Dito processo torna-se difícil pela grande quantidade e dispersão de dados. Assim, o Text Mining, apresenta-se como uma ferramenta alternativa que permitiria identificar as informações e categorizá-las de acordo ao seu potencial favorável ou desfavorável de influência, seja para uma organização ou para um setor empresarial.

3. OBJETIVOS

Objetivo geral: Fazer a classificação e o agrupamento de notícias usando diferentes técnicas, como Regressão, Redes Neurais e K-Means, aplicando seus resultados a diversas situações.

Objetivos específicos:

- Revisão bibliográfica de métodos de classificação e agrupamento aplicados em *Text Mining;*
- Criação de variáveis explicativas e variáveis resposta;
- Estudo do processo geral de modelagem;
- Estudo do *software* R para realizar a aplicação.

4. MATERIAIS E MÉTODOS

Na primeira parte do trabalho, será feito um levantamento bibliográfico com a leitura de artigos científicos, livros, matérias de jornal ou internet e vídeo aulas, a respeito de técnicas de categorização e agrupamento de dados. Para isto, serão utilizados computadores do *campus* com acesso à internet e a revistas científicas eletrônicas. Além disto, serão feitas pesquisas em livros da biblioteca.

Posteriormente, serão criadas variáveis explicativas e variáveis resposta para as diversas aplicações a serem realizadas neste projeto. As variáveis serão informações serão levantamentos de estatísticas como, por exemplo, indicador e frequência de determinados termos, quantidade de palavras, quantidade média de caracteres por palavra, percentual de conectivos, palavras mais frequentes e outras para buscar padrões que determinem cada tipo de notícia. Para esta etapa, é necessário computador com internet para a busca e tratamento das notícias a ser estudadas. Para a criação de variáveis, serão utilizados o *software R* e *SQL Server*.

Na etapa de modelagem, será utilizado o *software R*, que possui pacotes estatísticos para fazer modelagem e identificação de padrões. Para o adequado uso do R, serão utilizados manuais, tutoriais, artigos e páginas da internet sobre o *software*.

Para as análises dos resultados, serão utilizados principalmente os *softwares R* e *Excel.* Por último, para a confecção dos relatórios será usado o *Word.*

Este projeto é uma iniciativa do docente e não faz parte de um projeto maior, não contando com financiamento externo ou específico do IFSP.

Não há a necessidade da realização de viagens, a não ser para congressos que o aluno ou docente venha a participar para divulgar o trabalho.

Haverá a participação de professor pesquisador da PUC-Campinas em uma ou mais aplicações referentes à identificação de informações favoráveis/desfavoráveis a determinadas organizações ou setores empresariais. Além disto, mas terão discussões com outros docentes do *campus* além de apresentações e discussões do trabalho no grupo de pesquisa que o docente responsável participa.

5. PLANO DE TRABALHO

Tabela 5.1 Metas estabelecidas para a pesquisa.

METAS	DESCRIÇÃO
1	Revisão bibliográfica de técnicas para classificação e agrupamento de dados (Regressão, Redes Neurais, K-Means, etc.) para Text Mining
2	Estudo do processo de modelagem, como a separação em bases de desenvolvimento e validação, formas de evitar o *overfitting*, estatísticas para comparação de modelos.
3	Criação de variáveis explicativas, variáveis resposta e padronização dos dados.
4	Confecção e entrega do relatório parcial (até 06/07/18)
5	Implantação da aplicação e algoritmos
6	Análises dos resultados
7	Confecção e entrega do relatório final (entrega 30/11/2018)

Tabela 5.2 Cronograma proposta para cumprimento das metas.

	MESES								
METAS	MAR	ABR	MAI	JUN	JUL	AGO	SET	OUT	NOV
1	X	X	X	X					
2		X	X	X					
3		X	X	X					
4					X				
5					X	X	X		
6							X	X	
7								X	X

6. VIABILIDADE DE EXECUÇÃO

Os recursos necessários à pesquisa são computador com *softwares* instalados (*Word*, *Excel, SQL Server* e *R)* e artigos científicos, livros e outros materiais encontrados na internet. Além disto, o estudante fará cursos gratuitos na plataforma *Coursera* sobre os temas classificação, agrupamento de dados e modelagem em Mineração de Texto. Os textos para as aplicações foram obtidos e tratados no primeiro ano do projeto e são dos sites de notícia Diário do Grande ABC e Jornal do Brasil. Portanto, os dados já estão disponíveis para uso.

A pesquisa será realizada nas dependências do IFSP – Cubatão, principalmente em laboratórios de informática para a busca de artigos científicos, programação do algoritmo, análise dos resultados, confecção dos relatórios, reuniões entre orientador e aluno bolsista, e entre orientador/aluno com o pesquisador colaborador através de visitas pessoais ou de vídeo conferência. Pode-se também utilizar outras salas para etapas que não exijam o uso de computador, como a leitura de livros e artigos ou reuniões de discussão.

Para a realização deste projeto, é imprescindível a participação do bolsista e do corpo docente do *campus,* do professor pesquisador colaborador da PUC-Campinas, e do grupo de pesquisa (em que o docente responsável participa) nas

discussões das aplicações, além do corpo técnico-administrativo para auxiliar em questões específicas como o uso de salas, instalação de *softwares* e demais questões técnicas.

7. RESULTADOS ESPERADOS E DISSEMINAÇÃO

Neste projeto, serão aplicadas técnicas de modelagem em dados de notícias e espera-se, com isto, obter resultados satisfatórios na projeção, classificação e agrupamento destinados às aplicações do projeto, que são a projeção do preço das ações de empresas a partir da positividade/negatividade de notícias; a análise de tendência de jornais a respeito sobre determinados temas; a classificação de textos em categorias; e a identificação de informações e categorização de acordo ao seu potencial favorável/desfavorável de influência para uma organização ou setor empresarial. Além disto, também como como produto do projeto, terão os relatórios parcial e final do projeto, apresentações, artigos, resumos e outros trabalhos em eventos internos e externos, como congressos de iniciação científica.

Existe o potencial de inovação das aplicações, uma vez que, apesar do assunto *Text Mining* ser estudado por diversos pesquisadores, é um desafio extrair informações relevantes de um conjunto bruto de dados não estruturados. O sucesso das aplicações depende principalmente da criação de variáveis explicativas que podem ser as mais diversas possíveis, ficando sob a responsabilidade dos pesquisadores e sua liberdade criativa a busca por novas formas de extração de informações. No passo seguinte, de modelagem, o poder de classificação de um algoritmo espelha o poder explicativo das variáveis criadas anteriormente. Além destes fatores, não foram encontrados trabalhos na bibliografia que tenha aplicado *Text Mining* para identificar informações de acordo ao seu potencial favorável/desfavorável de influência para empresas ou setores.

A divulgação do trabalho será feita mediante apresentações, artigos, resumos, pôsteres ou outros possíveis meios em congressos internos do *campus* ou externos.

REFERÊNCIAS BIBLIOGRÁFICAS

AHLEMEYER-STUBBE, Andrea; COLEMAN, Shirley. **A Practical Guide to Data Mining for Business and Industry.** John Wiley & Sons, Ltd, 2014.

BARNEY, J. B.; HESTERLY, W. S. **Administração estratégica e vantagem competitiva: casos brasileiros**. 3ª Ed. São Paulo: Pearson Prentice Hall, 2011.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The Elements of Statistical Learning:** Data Mining, Inference and Predition. 2 ed, New York, Springer New York Inc, 2009.

IRELAND, R. Duane; HOSKISSON, Robert E; HITT, Michael A. **Administração Estratégica.** 10 ed. São Paulo: Cengage Learning, 2014.

KAO, Anne; POTEET, Stephen R.. **Natural Language Processing and Text Mining.** London, Springer-Verlag London, 2007.

ROGERS, Simon; GIROLAMI, Mark. **A First Course in Machine Learning.** CRC Press, 2011.

WEISS, Sholom M.; INDURKHYA, Nitin; ZHANG, Tong. **Fundamental of PredictiveText Mining.** 2 ed. London, Springer-Verlag London, 2015.