ISSN: 2316-6517

International Journal of Knowledge Engineering and Management

v. 09, n. 23, 2020.

ijkem.ufsc.br

International Journal of Knowledge Engineering and Management,
Florianópolis, v. 09, n. 23, p. 04-27, 2020.
• ISSN 2316-6517 •

A UTILIZAÇÃO DE CLASSIFICADORES DE TEXTO NA MINERAÇÃO DE IDEIAS AGREGANDO CRITÉRIOS DE ESPECIALISTAS

LUIZ FERNANDO SPILLERE DE SOUZA
Doutorando em Engenharia e Gestão do Conhecimento
Universidade Federal de Santa Catarina (UFSC)
spillere@gmail.com
ORCID: 0000-0002-0411-7252
ALEXANDRE LEOPOLDO GONÇALVES
Doutor em Engenharia de Produção
Universidade Federal de Santa Catarina (UFSC)
a.l.goncalves@ufsc.br
ORCID: 0000-0002-6583-2807

Submissão: 31 agosto 2019. Aceitação: 19 junho 2020.
Sistema de avaliação: duplo cego (*double blind review*).
UNIVERSIDADE FEDERAL DE SANTA CATARINA (UFSC)

International Journal of Knowledge Engineering and Management,
Florianópolis, v. 09, n. 23, p. 04-27, 2020.
• ISSN 2316-6517 •

A UTILIZAÇÃO DE CLASSIFICADORES DE TEXTO NA MINERAÇÃO DE IDEIAS AGREGANDO CRITÉRIOS DE ESPECIALISTAS

Resumo

Objetivo: A tarefa de buscar uma ideia em uma base de dados quando realizada por um especialista humano consiste em fazer a leitura de cada texto e estabelecer critérios de escolha para classificar textos que possam conter ou não ideias. O objetivo deste trabalho é aplicar a classificação de texto na mineração de ideias com a finalidade de simular a atuação de um especialista humano, comparando o desempenho dos classificadores de texto Support Vector Machines, Naive Bayes e Decision Trees.

Design | Metodologia | Abordagem: O método científico utilizado neste trabalho caracteriza-se como indutivo. A abordagem é quantitativa e a natureza refere-se a uma pesquisa aplicada. Quanto aos objetivos, a pesquisa é exploratória por proporcionar maior familiaridade com o problema, com vistas a torná-lo mais explícito ou construir hipóteses.

Resultados: A partir dos experimentos práticos realizados conclui-se que os classificadores de textos analisados apresentaram bom desempenho na tarefa de separar uma base de ideias segundo critérios utilizados por especialistas, através do aprendizado de máquina. A acurácia dos classificadores considerando individualmente critérios utilizados por especialistas variou entre 0,8014 a 0,9296, enquanto que o coeficiente kappa variou entre 0,5043 e 0,8419, superando as execuções onde não foram considerados os critérios.

Originalidade | Valor: A contribuição deste trabalho está fundamentada em agregar conhecimento de especialistas aos sistemas de aprendizado de máquina, mais especificamente aos classificadores de texto aplicados à mineração de ideias.

Palavra-chave: Classificação de Texto. Mineração de Ideias. Critérios de Especialistas.

International Journal of Knowledge Engineering and Management,
Florianópolis, v. 09, n. 23, p. 04-27, 2020.
• ISSN 2316-6517 •

THE USE OF TEXT CLASSIFIERS IN THE IDEA MINING ADDING EXPERT CRITERIA

Abstract

Goal: The task of searching for an idea in a database when performed by a human expert is to read each text and establish criteria of choice to classify texts that contain ideas or not. The aim of this paper is to apply text classification in idea mining in order to simulate the performance of a human expert, comparing the performance of the Support Vector Machines, Naive Bayes and Decision Trees text classifiers.

Design | Methodology | Approach: The scientific method used in this work is characterized as inductive. The approach is quantitative and nature refers to an applied research. Regarding the objectives, the research is exploratory once it provides greater familiarity with the problem aiming to make it more explicit or building hypotheses.

Results: From the practical experiments carried out, it can be concluded that the analyzed text classifiers performed well in the task of separating a database of ideas according to the criteria of an expert, through machine learning. The accuracy of the classifiers individually considering criteria used by experts ranged from 0.8014 to 0.9296, while the kappa coefficient varied between 0.5043 and 0.8419, surpassing the runs where such criteria were not considered.

Originality | Value: The contribution of this work is based on adding expert knowledge to machine learning systems, more specifically to text classifiers applied in the idea mining context.

Keywords: Text classification. Idea Mining. Expert Criteria.

International Journal of Knowledge Engineering and Management,
Florianópolis, v. 09, n. 23, p. 04-27, 2020.
• ISSN 2316-6517 •

1 Introdução

A tarefa de classificação de texto surgiu da necessidade natural de organização de dados em que se necessita agrupar assuntos semelhantes. Os métodos automáticos, por sua vez, vieram para auxiliar esta tarefa diante do aumento no volume de dados disponíveis para consulta.

Uma das definições de classificação de texto refere-se como o processo de classificar documentos de texto em um número fixo de classes predefinidas (VIJAYAN; BINDU; PARAMESWARAN, 2017). De forma semelhante, Altinel e Ganiz (2018), definem a classificação automática de texto como a tarefa de organizar documentos em classes predeterminadas, geralmente usando algoritmos de aprendizado de máquina.

A classificação de texto define que os objetos são separados em categorias, geralmente para algum propósito específico, onde uma categoria explora uma relação entre as palavras e os seus significados. É uma tecnologia chave para lidar e organizar grandes volumes de documentos, sendo utilizada em aplicações de gerenciamento de informações, alocando automaticamente um documento para uma ou mais classes predefinidas (KADHIM, 2019).

A tarefa de classificação de texto encontra aplicações em uma ampla variedade de domínios na mineração de texto, dentre os quais podem ser citados: organização e filtragem de notícias, organização e recuperação de documentos, mineração de ideias, classificação de e-mail e filtragem de spam (AGGARWAL; ZHAI, 2012). No contexto de mineração de ideias existem trabalhos publicados comprovando a sua eficiência (ALKSHER et al., 2018a, 2018b; CHRISTENSEN et al., 2017a; LIU; GOULDING; BRAILSFORD, 2015; THORLEUCHTER; VAN DEN POEL; PRINZIE, 2010).

O objetivo deste trabalho é aplicar a classificação de texto à mineração de ideias com a finalidade de simular a atuação de um especialista humano, comparando o desempenho dos classificadores de texto Support Vector Machines, Naive Bayes e Decision Trees. Resumidamente, ao buscar uma ideia em uma base de dados, o especialista humano realiza uma leitura estabelecendo alguns critérios de escolha para a seleção do texto. A partir desta separação os textos enquadrados nos critérios dos especialistas são eleitos como ideias. Para Keller e Kotler (2012), a seleção de ideias pode ser vista como um processo onde se avaliam aquelas que atenderem aos critérios, classificando-as por meio de um método a ser escolhido.

International Journal of Knowledge Engineering and Management,
Florianópolis, v. 09, n. 23, p. 04-27, 2020.
• ISSN 2316-6517 •

Este trabalho está estruturado em cinco seções: a primeira, intitulada introdução, faz uma breve definição da classificação de texto e suas aplicações, bem como apresenta os objetivos da pesquisa realizada. A segunda seção contém uma contextualização da classificação de texto aplicada à mineração de ideias, com uma fundamentação teórica. A seção 3 apresenta os procedimentos metodológicos utilizados e a seção 4 apresenta uma análise e discussão dos resultados obtidos. Por fim, na seção 5 são expostas as considerações obtidas ao final do estudo.

2 Classificação de texto aplicada à mineração de ideias

Desde o início do uso do termo mineração de ideias, por volta dos anos de 2010, sua utilização vem sendo combinada com a classificação de texto. Estudos iniciais como os de Thorleuchter, Van Den Poel e Prinzie (2010) já utilizavam a classificação de texto (tokenização, métodos de filtragem de termos, medida de distância euclidiana, entre outros) combinada com medidas heurísticas para mineração de ideias.

A partir de então a mineração de ideias evoluiu paralelamente aos métodos de classificação de texto, ganhando robustez e confiabilidade. Dentre os trabalhos mais atuais cita-se Alksher et al. (2018a), em que os autores demonstram que é possível detectar ideias a partir de dados gerados em comunidades online automaticamente, utilizando classificadores baseados em aprendizagem de máquina (neste caso Support Vector Machines).

O presente trabalho propõe a utilização de classificadores de textos de cunho geral como Support Vector Machines, Decision Trees e Naive Bayes simulando os critérios utilizados por especialistas humanos na tarefa de localizar possíveis ideias a partir de textos. Os testes realizados demonstram que a partir de uma base de dados o classificador é capaz de identificar textos a partir dos critérios pré-estabelecidos.

3 Fundamentação teórica

A classificação de texto é usada para extrair conhecimento a partir de padrões de texto não estruturado. É uma área de pesquisa que assume o desafio de produzir ferramentas de inteligência, analisar grandes quantidades de texto em linguagem natural e encontrar padrões (BRINDHA; PRABHA; SUKUMARAN, 2016).

O processo de classificação de texto é algo complexo, até mesmo para especialistas de domínio pelo fato destes utilizarem critérios subjetivos, o que pode gerar divergência entre as classificações de um mesmo texto, quando realizado por mais de um especialista humano (CECI; WOSZEZENKI; GONÇALVES, 2014).

Uma das formas de contornar isto é realizar a classificação de texto de maneira automática utilizando abordagens baseadas em regras e abordagens de aprendizado de máquina. Na abordagem baseada em regras, as regras de classificação são definidas através de programação e os documentos são classificados com base nestas regras. A abordagem baseada em regras promove bons resultados quando o número de regras é pequeno, caso contrário, a manutenção da base de regras se torna difícil à medida que o número de regras aumenta e acabam conflitando entre si (SEBASTIANI, 2002).

Para superar essas limitações, a abordagem de aprendizado de máquina é usada na classificação de texto. Ele classifica documentos de texto observando as características de um conjunto de documentos e, a partir dessas características, o classificador tem a tarefa de decidir para qual categoria um novo documento desconhecido será atribuído (DWIVEDI; ARYA, 2016).

3.1 Support Vector Machines

Uma máquina de vetores de suporte (Support Vector Machine - SVM) tem como princípio determinar um limiar de separação em um espaço de busca que pode separar melhor diferentes classes (AGGARWAL; ZHAI, 2012). Matematicamente, uma SVM gera uma decisão limite que melhor separa duas classes com uma margem em torno deste limite, que tem sua posição e largura controlada e pode ser visto como um parâmetro utilizado para ajustar a sensibilidade do classificador (CHRISTENSEN et al., 2017b).

Com sua boa capacidade de generalização, as SVMs eliminam a necessidade de seleção de características, tornando a aplicação da categorização do texto consideravelmente mais fácil. Outra vantagem das SVMs sobre os métodos convencionais é a sua robustez. SVMs mostram um bom desempenho em diversos cenários de experimentos, evitando falhas como observado em alguns métodos tradicionais. Além disso, as SVMs não exigem nenhum parâmetro de ajuste, pois eles

podem encontrar boas configurações de parâmetros automaticamente (JOACHIMS, 1998).

Estudos como o de Coussement e Van den Poel (2008), mostram que as SVMs apresentam bom desempenho de generalização quando aplicados à mineração de dados utilizando grandes bases de dados e com muitos ruídos. Já Christensen et.al. (2017) usaram uma SVM para testar se um classificador de aprendizado de máquina desta natureza poderia aprender o padrão de ideias escritas como texto. A comparação entre o desempenho no conjunto de validação e o desempenho no conjunto de teste demonstra a confiabilidade do classificador utilizado.

3.2 Decision Trees

Uma árvore de decisão (Decision Tree - DT) é essencialmente uma decomposição hierárquica do espaço de dados (treinamento), em que um predicado ou uma condição no valor do atributo são usados para dividir o espaço de dados hierarquicamente. A divisão do espaço de dados é realizada recursivamente na árvore de decisão, até que os nós da folha contenham certo número mínimo de registros (QUINLAN, 1986).

O documento é então classificado para a classe representada pelo nó da folha. Os predicados de decisão nos nós internos podem ser a presença ou ausência dos termos em documentos de texto (VIJAYAN; BINDU; PARAMESWARAN, 2017).

A construção de uma árvore de decisão consiste em partições sucessivas do conjunto de treinamento original em subconjuntos menores. Em um contexto de mineração de dados, mesmo que não sejam necessariamente utilizadas na classificação de novas instâncias, as DTs podem ser construídas para fornecer descrições das características comuns aos membros de cada classe (FRIZZARINI; LAURETTO, 2013).

Os classificadores de árvore de decisão mostram um grande potencial em muitos problemas de reconhecimento de padrões, tais como: classificação de dados de várias origens, diagnóstico médico, reconhecimento de fala, entre outros. Uma das principais características das DTs reside na flexibilidade de serem utilizadas com diferentes subconjuntos de recursos e regras de decisão, em diferentes estágios de classificação, bem como a capacidade de compensações entre precisão de classificação e eficiência de tempo (SAFAVIAN; LANDGREBE, 1991).

International Journal of Knowledge Engineering and Management,
Florianópolis, v. 09, n. 23, p. 04-27, 2020.
• ISSN 2316-6517 •

3.3 Naive Bayes

Os classificadores Naive Bayes (NB) adotam a suposição de que o valor de uma determinada característica é independente do valor de qualquer outra característica em um texto. Na classificação de texto, a suposição em um classificador NB é que a probabilidade de cada palavra que aparece em um documento é independente da ocorrência de outra palavra no mesmo documento (DENG et al., 2019).

Em Zhang e L1 (2007) os autores utilizaram NB para realizar a detecção de spam e mencionam a necessidade de um número elevado de instâncias de treinamento para uma classificação precisa. O mesmo estudo sugere um ajuste dinâmico das probabilidades de ocorrência das palavras (características) durante a classificação para obter um modelo capaz de realizar previsões adequadas.

O NB é um algoritmo clássico e tem sido amplamente utilizado na categorização de textos (JOACHIMS, 1997; MCCALLUM; NIGAM, 1998). O classificador NB simplifica muito o aprendizado e compete bem com classificadores mais sofisticados, sendo indicado para situações onde é necessário escolher entre duas condições distintas de uma classe de parâmetros, ou seja, escolhas binárias (RISH, 2001). Seu desempenho competitivo na classificação é surpreendente, uma vez que a suposição inicial de independência condicional em que se baseia normalmente é verdadeira em uma série de aplicações do mundo real (ZHANG, 2004).

Embora o algoritmo NB seja caracterizado pela simplicidade, trabalhos recentes como Ababneh (2019) e Chen et al. (2019) demonstraram uma boa eficiência na tarefa de classificação de texto em mais de um idioma e com conjuntos de treinamentos pequenos.

4 Procedimentos metodológicos

Esta pesquisa tem por objetivo gerar conhecimento para aplicações em mineração de ideias utilizando classificadores de texto para simular critérios de especialistas na escolha de ideias.

O método científico utilizado segue o método indutivo permitindo que o pesquisador defina uma hipótese a respeito de um objeto de valor científico e que, sendo confirmada pela experimentação controlada, permite que os resultados sejam

generalizados sob a forma de método, lei ou teoria (LAKATOS; MARCONI, 2010). A abordagem é quantitativa, pois os resultados da pesquisa podem ser quantificados, recorrendo à linguagem matemática para descrever as causas de um fenômeno e as relações entre variáveis (FONSECA, 2002). Quanto à natureza é uma pesquisa aplicada, pois objetiva gerar conhecimentos para aplicação prática e dirigidos à solução de problemas específicos (SILVA; MENEZES, 2005). E por último, quanto aos objetivos, é exploratória, pois examina um conjunto de fenômenos buscando anomalias que não sejam ainda conhecidas e, que possam ser então a base para uma pesquisa mais elaborada (WAZLAWICK, 2010).

Este trabalho apresenta o resultado de um estudo prático, sendo que para isto foram definidas as seguintes etapas:

1) Obtenção de uma base de dados de ideias já classificada por especialistas;

2) Inclusão de textos comuns nesta base de dados de ideias;

3) Definição de três critérios utilizados por especialistas em seleção de ideias baseados em pesquisas na literatura, sendo eles: originalidade, produtibilidade e viabilidade econômica;

4) Separação da base de dados de ideias onde apareçam os critérios definidos no item anterior;

5) Submissão da base de dados aos classificadores de texto Naive Bayes, Support Vector Machines e Decision Trees para testar a capacidade dos classificadores em separar textos que representem ideias de textos comuns;

6) Nova submissão da base de dados aos classificadores Naive Bayes, Support Vector Machines e Decision Trees para testar a capacidade dos classificadores em separar ideias a partir dos 3 critérios definidos no item 3 dos textos comuns

7) Nova Submissão da base de dados aos classificadores Naive Bayes, Support Vector Machines e Decision Trees testar a capacidade dos classificadores em separar cada critério individual dos especialistas definidos no item 3 dos textos comuns;

8) Comparação dos resultados obtidos (acurácia e kappa) sobre base de dados de ideias considerando os itens 5, 6 e 7.

A base de dados utilizada neste estudo contém ideias disponibilizadas publicamente através do Portal Sinapse da Inovação®, que é um programa de incentivo ao empreendedorismo inovador que tem por objetivo “transformar e aplicar as boas ideias geradas por estudantes, pesquisadores, professores e profissionais dos

International Journal of Knowledge Engineering and Management,
Florianópolis, v. 09, n. 23, p. 04-27, 2020.
• ISSN 2316-6517 •

diferentes setores do conhecimento e econômicos em negócios de sucesso" (SINAPSE, 2017). O conjunto de dados possui 122 textos representando ideias que alcançaram a última etapa para serem selecionadas, sendo aprovadas e que receberam aporte financeiro do Sinapse na Inovação®. Adicionalmente, foram incluídos ao conjunto de dados 100 textos com conteúdos diversos retirados da web.

O conjunto de dados de ideias foi então classificado manualmente, a partir de uma leitura, onde os números foram: ideias que explicitamente mencionam a viabilidade econômica (rótulo Viabilidade Econômica), 39; ideias que mencionam informações que possam representar o critério de produtibilidade (rótulo Produtibilidade), 42 e ideias que em seu texto denotam um grau elevado de originalidade (rótulo Originalidade), 41. Por fim, foram rotulados os 100 textos comuns extraídos aleatoriamente da web (rótulo Texto Comum) mostrados na Figura 1.

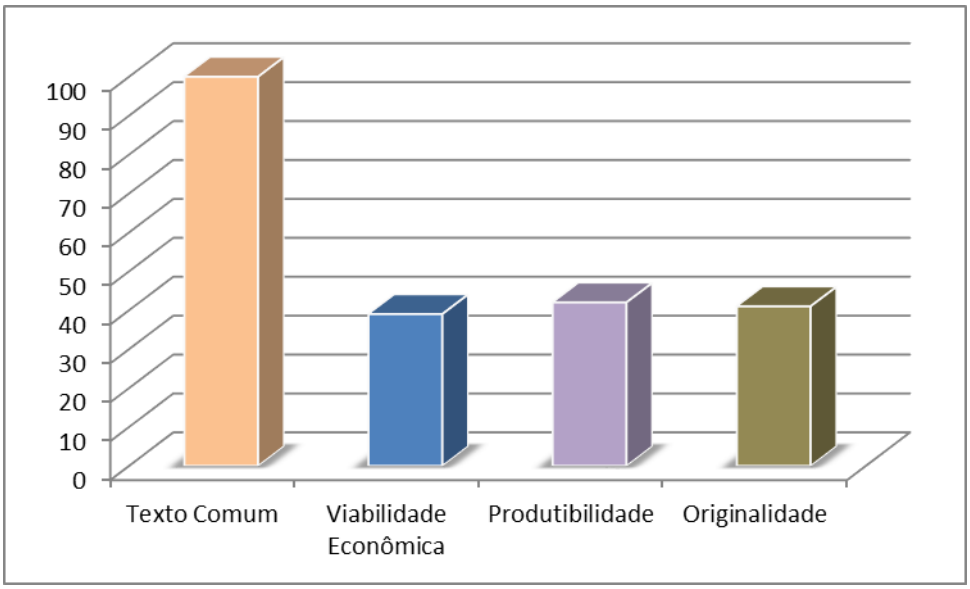

Figura 1 – Separação do conjunto dados. Fonte: Elaborado pelos autores (2019).

Para a análise do conjunto de dados utilizou-se a aplicação Lightside®, por ser um aplicativo de código aberto destinado à tarefa de classificação de textos. Foram utilizadas as configurações padrões do classificador SVM, habilitando a opção LibLinear, que proporciona um pareamento de classe e rótulos, adequada para aumentar a eficiência do classificador. Também foram utilizadas as configurações padrões do classificador DT, desabilitando as opções de trabalho com tabelas esparsas e de alta dimensão, condição mais indicada para trabalhos que envolvem valores

numéricos. E por último, as configurações padrões do classificador NB foram utilizadas, onde se encontram desabilitadas as opções de utilização de estimadores Kernel e discretização supervisionada.

Para a configuração dos recursos foi utilizada a opção de unigrams onde são extraídas as palavras distintas do conteúdo de cada texto do conjunto de dados. Cada palavra representa uma característica, tendo agregada a si a frequência em que ocorre na coleção de textos. Também foi utilizada a opção include pontuaction indicando que a pontuação é incluída na análise das palavras, pois a pontuação pode ser uma fonte crucial de informação para algumas tarefas, assim como a opção track feature hit location que inclui o registro da localização de cada característica extraída de cada documento. Isso é importante para permitir que se execute a análise de erros depois de ter criado e avaliado um modelo (MAYFIELD; ROSÉ, 2012).

Após a execução utilizando os algoritmos de SVM, NB e DT, foram analisadas as matrizes de confusão e as métricas de avaliação de acurácia e o coeficiente kappa. A matriz de confusão, também conhecida como matriz de erro, representa um layout de tabela específico que permite a visualização do desempenho de um algoritmo, tipicamente voltado ao aprendizado supervisionado, mostrando as classificações corretas e incorretas. A acurácia é uma métrica simples que informa o desempenho de um classificador ao prever rótulos de classe a partir de seu conjunto de dados. Já o coeficiente kappa mede o nível de concordância entre dois conjuntos de dados (LANDIS; KOCH, 1977).

Em complemento à etapa 3 e considerando a revisão da literatura, foram identificados vários critérios de especialistas que são levados em consideração em sua tarefa de selecionar ideias. Para a realização dos testes, foram escolhidos três destes critérios, listados e referenciados a seguir:

- Originalidade: representa a novidade e a inovação, isto é, quão incomum e original as ideias são a respeito do contexto em que se inserem (MAGNUSSON; WÄSTLUND; NETZ, 2014). O autor baseia-se em Amabile (1996), no qual defende que a originalidade é um conceito genérico e que as pessoas têm uma sensação intuitiva para o que é criativo;
- Produtibilidade: representa a perspectiva da empresa sobre a facilidade com que o serviço pode ser implementado e/ou produzido. Este critério assume a perspectiva da oferta (MAGNUSSON; WÄSTLUND; NETZ, 2014);

International Journal of Knowledge Engineering and Management,
Florianópolis, v. 09, n. 23, p. 04-27, 2020.
• ISSN 2316-6517 •

• Viabilidade econômica: para aprofundar o caráter econômico durante a avaliação da ideia, é essencial estimar o preço de venda e o benefício potencial. Também é importante avaliar se está de acordo com os objetivos da empresa, estimar os custos de produção e de desenvolvimento, bem como o tempo que vai levar para atingir o mercado (OZER, 2004; FERIOLI et al., 2008).

5 Análise e discussão dos resultados

A primeira execução dos classificadores NB, SVM e DT levou em consideração o conjunto de dados produzido para este estudo. Teve por finalidade a verificação do desempenho dos classificadores na separação simples dos textos contendo ideias dos textos contendo texto comum. A seguir são apresentadas as matrizes de confusão (conforme Tabela 1) geradas a partir desta execução.

Tabela 1 – Matriz de confusão considerando a separação de ideias dos textos comuns.

Model Confusion Matrix: Naive Bayes		
Act \ Pred	Ideias	Texto Comum
Ideias	97	25
Texto Comum	21	79
Model Confusion Matrix: Support Vector Machines		
Act \ Pred	Ideias	Texto Comum
Ideas	115	7
Texto Comum	19	81
Model Confusion Matrix: Decision Trees		
Act \ Pred	Ideias	Texto Comum
Ideas	97	25
Texto Comum	21	79

Fonte: Elaborado pelos autores (2019).

A Tabela 1 mostra que os classificadores foram capazes de separar bem os textos comuns dos textos que representam ideias. A matriz de confusão assinala em azul os acertos realizados pelos classificadores.

International Journal of Knowledge Engineering and Management,
Florianópolis, v. 09, n. 23, p. 04-27, 2020.
• ISSN 2316-6517 •

A segunda execução dos classificadores NB, SVM e DT também levou em consideração o conjunto de dados produzido para este estudo, porém analisando a capacidade dos classificadores em separar as ideias conforme os três critérios estabelecidos dos textos contendo texto comum, conforme representadas no Gráfico 1. A seguir são apresentadas as matrizes de confusão (conforme Tabela 2) geradas a partir desta execução.

Tabela 2 – Matriz de confusão considerando as quatro classes.

Naive Bayes				
Act \ Pred	Originalidade	Produtibilidade	Viabilidade Econômica	Texto Comum
Originalidade	19	10	11	1
Produtibilidade	10	21	11	0
Viabilidade Econômica	12	13	14	0
Texto Comum	0	0	0	100
Support Vector Machines				
Act \ Pred	Originalidade	Produtibilidade	Viabilidade Econômica	Texto Comum
Originalidade	17	11	10	3
Produtibilidade	10	19	13	0
Viabilidade Econômica	9	14	16	0
Texto Comum	0	0	0	100
Decision Trees				
Act \ Pred	Originalidade	Produtibilidade	Viabilidade Econômica	Texto Comum
Originalidade	22	5	13	1
Produtibilidade	10	15	17	0
Viabilidade Econômica	11	11	12	5
Texto Comum	0	4	4	92

Fonte: Elaborado pelos autores (2019).

International Journal of Knowledge Engineering and Management,
Florianópolis, v. 09, n. 23, p. 04-27, 2020.
• ISSN 2316-6517 •

A Tabela 2 mostra que os classificadores foram capazes de separar bem os textos comuns dos textos que representam ideias, de forma semelhante à primeira execução. Todavia, quando analisados os critérios originalidade, produtibilidade e viabilidade econômica, percebe-se uma baixa capacidade em identificar corretamente as classes em que as ideias foram divididas.

Em uma terceira execução, os classificadores NB, SVM e DT foram executados novamente considerando os textos comuns em relação aos textos de ideias referentes a cada critério individualmente. A finalidade deste teste foi verificar o desempenho dos classificadores nesta tarefa de separação em duas classes distintas: cada critério individualmente comparado aos textos comuns. Segue abaixo a matriz de confusão (conforme Tabela 3, 4 e 5) gerada a partir destas execuções.

Tabela 3 – Matriz de confusão considerando as ideias rotuladas como Originalidade versus Texto Comum.

Model Confusion Matrix: Naive Bayes		
Act \ Pred	Originalidade	Texto Comum
Originalidade	39	2
Texto Comum	10	90
Model Confusion Matrix: Support Vector Machines		
Act \ Pred	Ideias	Texto Comum
Originalidade	32	9
Texto Comum	9	91
Model Confusion Matrix: Decision Trees		
Act \ Pred	Ideias	Texto Comum
Originalidade	25	16
Texto Comum	12	88

Fonte: Elaborado pelos autores (2019).

International Journal of Knowledge Engineering and Management,
Florianópolis, v. 09, n. 23, p. 04-27, 2020.
• ISSN 2316-6517 •

Tabela 4 – Matriz de confusão considerando as ideias rotuladas como Produtibilidade versus Texto Comum.

Model Confusion Matrix: Naive Bayes		
Act \ Pred	Produtibilidade	Texto Comum
Produtibilidade	42	0
Texto Comum	10	90
Model Confusion Matrix: Support Vector Machines		
Act \ Pred	Produtibilidade	Texto Comum
Produtibilidade	39	3
Texto Comum	7	93
Model Confusion Matrix: Decision Trees		
Act \ Pred	Produtibilidade	Texto Comum
Produtibilidade	29	13
Texto Comum	10	90

Fonte: Elaborado pelos autores (2019).

Tabela 5 – Matriz de confusão considerando as ideias rotuladas como Viabilidade Econômica versus Texto Comum.

Model Confusion Matrix: Naive Bayes		
Act \ Pred	Viabilidade Econômica	Texto Comum
Viabilidade Econômica	39	0
Texto Comum	11	89
Model Confusion Matrix: Support Vector Machines		
Act \ Pred	Viabilidade Econômica	Texto Comum
Viabilidade Econômica	35	4
Texto Comum	7	93
Model Confusion Matrix: Decision Trees		
Act \ Pred	Viabilidade Econômica	Texto Comum
Viabilidade Econômica	27	12
Texto Comum	13	87

Fonte: Elaborado pelos autores (2019).

International Journal of Knowledge Engineering and Management,
Florianópolis, v. 09, n. 23, p. 04-27, 2020.
• ISSN 2316-6517 •

De acordo com Campbell e Wynne (2011), os erros estão presentes em qualquer tipo de classificação e a forma padronizada para reportar erros em locais específicos é a chamada Matriz de Confusão. Esta matriz identifica não somente o erro global da classificação para cada categoria, mas também como se deram as confusões entre categorias.

A partir da matriz de confusão da Tabela 1, verificou-se que na primeira execução ocorreram poucos erros de classificação entre ideias e textos comuns. Isto demonstra que os classificadores analisados foram capazes de separar as ideias dos textos comuns. Porém, a partir da matriz de confusão da Tabela 2, verificou-se que nesta segunda execução os classificadores tiveram queda de desempenho ao classificar os três critérios presentes nas ideias.

Para contornar esta deficiência foi realizada uma terceira execução separando novamente o conjunto de dados das ideias em critérios individuais e confrontando-os com texto comum, caracterizando então uma classificação binária. A matriz de confusão de cada critério (Tabelas 3, 4 e 5) demonstrou que neste caso, os classificadores conseguiram melhorar seu índice de acerto.

Existem na literatura alguns trabalhos, dentre os quais podem ser citados, Li, Liu e Ng (2010), Badawi e Altinçay (2014), Elhassan e Ali (2019) e em que os classificadores treinados para escolhas binárias apresentam melhor desempenho comparados com a utilização de mais de duas classes de escolha. A complexidade exigida de um classificador é maior conforme o aumento no número de classes (ELHASSAN; ALI, 2019).

Finalmente é apresentada a tabela contendo os índices de acurácia e o coeficiente kappa, onde os desempenhos de todas as execuções podem ser analisados e comparados entre si (conforme Tabela 6).

International Journal of Knowledge Engineering and Management,
Florianópolis, v. 09, n. 23, p. 04-27, 2020.
• ISSN 2316-6517 •

Tabela 6 – Índice de acurácia e coeficiente kappa dos algoritmos NB, SVM e DT nos diversos cenários

	Naive Bayes		*Support Vector Machines*		*Decision Trees*	
	Acurácia	*Kappa*	*Acurácia*	*Kappa*	*Acurácia*	*Kappa*
Ideias x Textos Comuns (2 classes) – 1ª execução	0,7928	0,5830	0,8829	0,7609	0,7928	0,5830
Ideias com os Critérios x Textos Comuns (4 classes) – 2ª execução	0,6937	0,5593	0,6847	0,5448	0,6351	0,4781
Critério de Viabilidade Econômica x Textos Comuns (2 classes) – 3ª execução	0,9209	0,8195	0,9209	0,8085	0,8201	0,5579
Critério de Produtividade x Textos Comuns (2 classes) – 3ª execução	0,9296	0,8419	0,9296	0,8355	0,8380	0,6030
Critério de Originalidade x Textos Comuns (2 classes) – 3ª execução	0,9149	0,8049	0,8723	0,6905	0,8014	0,5043

Fonte: Elaborado pelos autores (2019).

Os dados desta tabela demonstram que os classificadores Naive Bayes, Support Vector Machines e Decision Trees apresentam desempenho similar considerando a tarefa de separar os textos contendo ideias dos textos comuns (1ª execução), tendo o SVM atingido os melhores resultados. Quando os textos contendo ideias foram divididos em critérios (2ª execução), o desempenho dos classificadores Naive Bayes, Support Vector Machines e Decision Trees apresentam uma queda, que pode ser explicada em virtude do aumento no número de classes e, consequentemente, da complexidade. Quando os critérios foram considerados individualmente (3ª execução), os classificadores apresentam desempenho melhor se comparado às execuções anteriores. Cabe ressaltar que, quando os critérios são separados, os classificadores

Naive Bayes e Support Vector Machines tiveram desempenho superior ao classificador Decision Trees.

Outro ponto importante é a observação do coeficiente kappa, que também demonstra que os classificadores Naive Bayes e Support Vector Machines apresentam desempenho melhor quando os critérios foram avaliados individualmente (3ª execução). Nestes casos, o coeficiente kappa encontrado variou entre 0,5043 a 0,8419, indicando um desempenho entre bom e muito bom de acordo com Landis e Koch (1977).

A partir desta análise é possível verificar que a melhor abordagem consiste na realização da classificação de textos utilizando cada critério de um especialista de forma individual, sendo que os classificadores Nayve Bayes e Support Vector Machines tiveram o melhor desempenho ao realizarem esta tarefa.

Cabe ressaltar que o conjunto de dados utilizado no treinamento dos classificadores é pequeno e homogêneo, o que pode ser um dos motivos para os elevados índices atingidos. Sabe-se que em condições reais, os dados podem ser esparsos e incompletos, gerando assim uma queda neste desempenho. Porém, a constatação de que os classificadores Naive Bayes e Support Vector Machines e Decision Trees são capazes de separar adequadamente textos que representam ideias de textos comuns fica comprovada pelos testes realizados.

6 Considerações finais

O estudo realizado testou os classificadores de textos Support Vector Machines, Decision Trees e Naive Bayes aplicados à mineração de ideias simulando os critérios utilizados por especialistas humanos na tarefa de buscar ideias. Os resultados demonstram que, a partir de um conjunto de dados, os classificadores apresentaram um desempenho melhor ao serem treinados para buscar um critério individual utilizado por um especialista versus texto comum (terceira execução). Quando estes critérios não estão separados, o desempenho destes classificadores é menor (primeira e segunda execução). Esta constatação contribui para estudos posteriores em mineração de ideias realizarem buscas de ideias a partir de critérios específicos.

Portanto, a principal contribuição deste estudo é mostrar que os classificadores são capazes de agregar à tarefa de mineração de ideias a capacidade que, a partir de critérios pré-estabelecidos, um especialista humano possui de separar ideias. Através

de técnicas de aprendizado de máquina e classificação de texto foi possível simular como um especialista humano faria, através de critérios, a classificação de um texto como sendo uma ideia.

Como trabalhos futuros, pretende-se testar outros classificadores baseados em aprendizado de máquina. Também se mostra interessante a utilização de outros conjuntos de dados, se possível com uma quantidade maior de textos. Todavia, existe a dificuldade de encontrar bases de dados abertas que abordem inovação, devido ao fato de representarem conhecimento estratégico nas empresas. Por último, objetiva-se agregar outros critérios não listados neste trabalho que também são mencionados na literatura como critérios de escolha de ideias.

7 Referências bibliográficas

ABABNEH, J. Application of Naïve Bayes, Decision Tree, and K-Nearest Neighbors for Automated Text Classification. **Modern Applied Science**, v. 13, n. 11, p. 31, 2019.

AGGARWAL, C. C.; ZHAI, C. X. A survey of text classification algorithms. **Mining Text Data**, v. 9781461432, p. 163–222, 2012.

ALKSHER, M.; AZMAN, A.; YAAKOB, R.; KADIR, R. A. Feasibility of Using the Position as Feature for Idea Identification from Text. 2018 **Fourth International Conference on Information Retrieval and Knowledge Management** (CAMP), p. 1–6, 2018a.

ALKSHER, M.; AZMAN, A.; YAAKOB, R.; ALSHARI, E. Effective idea mining technique based on modeling lexical semantic. **Journal of Theoretical and Applied Information Technology**, v. 96, n. 16, p. 5350-5362, 2018.

ALTINEL, B.; GANIZ, M. C. Semantic text classification: A survey of past and recent advances. **Information Processing and Management**, v. 54, n. 6, p. 1129–1153, 2018.

AMABILE, T. M. **Creativity in context:** The social psychology of creativity Boulder, CO: Westview, 1996.

BADAWI, D.; ALTINÇAY, H. A novel framework for termset selection and weighting in binary text classification. **Engineering Applications of Artificial Intelligence**, v. 35, p. 38–53, 2014.

BRINDHA, S.; PRABHA, K.; SUKUMARAN, S. A survey on classification techniques for text mining. In: INTERNATIONAL CONFERENCE ON ADVANCED COMPUTING AND COMMUNICATION SYSTEMS: Bringing to the Table, Futuristic Technologies from Arround the Globe. **Anais…** ICACCS 2016, v. 01, n. i, p. 1–5, 2016.

CAMPBELL, J. B.; WYNNE, R. H. **Introduction to remote sensing Guilford Press**, 2011.

CECI, F.; WOSZEZENKI, C. R.; GONÇALVES, A. L. O uso de anotações semânticas e ontologias para a classificação de documentos. **International Journal of Knowledge Engineering and Management (IJKEM)**, v. 3, n. 5, p. 1-14, 2014.

CHEN, J.; DAI, Z.; DUAN, J.; MATZINGER, H.; POPESCU, L. Naive bayes with correlation factor for text classification problem. In: IEEE INTERNATIONAL CONFERENCE ON MACHINE LEARNING AND APPLICATIONS, **Anais…** ICMLA 2019, p. 1051–1056, 2019.

CHRISTENSEN, K.; NORSKOV, S.; FREDERILSEN, L.; SHOLDERER, J. In Search of New Product Ideas: Identifying Ideas in Online Communities by Machine Learning and Text Mining. **Creativity and Innovation Management**, v. 26, n. 1, p. 17–30, 2017a.

CHRISTENSEN, K.; LILAND, K. H.; KVAAL, K.; RISVIK, E.; BIANCOLILLO, A.; SHOLDERER, J.; NORSKOV, S.; NAES,T. Mining online community data: The nature of ideas in online communities. **Food Quality and Preference**, v. 62, n. December 2016, p. 246–256, 2017b.

COUSSEMENT, K.; VAN DEN POEL, D. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. **Expert Systems with Applications**, v. 34, n. 1, p. 313–327, 2008.

DENG, X. et al. **Feature Selection for Text Classification**. p. 257–276, 2019.

DWIVEDI, S. K.; ARYA, C. **Automatic Text Classification in Information retrieval**. p. 1–6, 2016.

ELHASSAN, R.; ALI, M. The Impact of Feature Selection Methods for Classifying Arabic Texts. In: INTERNATIONAL CONFERENCE ON COMPUTER APPLICATIONS AND INFORMATION SECURITY, **Anais…** ICCAIS 2019, 2019.

FERIOLI, M. et al. Evaluation of the Potential Performance of Innovative Concepts in the Early Stages of the New-Product Development Process (Npdp). **Design**, p. 1139–1148, 2008.

FONSECA, J. J. S. DA. **Metodologia da pesquisa científica**. São Carlos: Serviço de Biblioteca e Informação, 2002.

FRIZZARINI, C.; LAURETTO, M. S. Proposta de um Algoritmo para Indução de Árvores de Classificação para Dados Desbalanceados. In: SIMPÓSIO BRASILEIRO DE SISTEMAS DE INFORMAÇÃO, **Anais...** v. IX, p. 722–733, 2013.

JOACHIMS, T. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. **Fourteenth International Conference on Machine Learning**, p. 143–151, 1997.

JOACHIMS, T. Text categorization with support vector machines: Learning with many relevant features. **Lecture Notes in Computer Science** (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), v. 1398, p. 137–142, 1998.

KADHIM, A. I. Survey on supervised machine learning techniques for automatic text classification. **Artificial Intelligence Review**, v. 52, n. 1, p. 273–292, 2019.

KELLER, K. L.; KOTLER, P. **Administração de Marketing**. 12. ed. São Paulo: [s.n.].

LAKATOS, E. M.; MARCONI, M. DE A. **Metodologia científica**. 2. ed. São Paulo: Atlas, 2010.

LANDIS, J. R.; KOCH, G. G. The Measurement of Observer Agreement for Categorical Data. **Biometrics**, v. 33, n. 1, p. 159, 1977.

LI, X. L.; LIU, B.; NG, S. K. Negative training data can be harmful to text classification. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, PROCEEDINGS OF THE CONFERENCE. **Anais…** EMNLP 2010. n. October, p. 218–228, 2010.

LIU, H.; GOULDING, J.; BRAILSFORD, T. Towards computation of novel ideas from corpora of scientific text. **Lecture Notes in Computer Science** (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), v. 9285, p. 541–556, 2015.

MAGNUSSON, P. R.; WÄSTLUND, E.; NETZ, J. Exploring Users' Appropriateness as a Proxy for Experts When Screening New Product/Service Ideas. **Journal of Product Innovation Management**, v. 33, n. 1, p. 4–18, 2014.

MAYFIELD, E.; ROSÉ, C. P. **LightSIDE:** Text Mining and Machine Learning User' s Manual. [s.l: s.n.].

MCCALLUM, A.; NIGAM, K. **Employing EM and Pool-Based Active Learning for Text Classi cation 2 Probabilistic Framework for Text Classi cation**. 1998.

OZER, M. Managing the selection process for new product ideas. **Research Technology Management**, v. 47, n. 4, p. 11, 2004.

QUINLAN, J. R. Induction of Decision Trees. **Research and Development in Expert Systems XV**, v. 1, n. Chapter 2, p. 15–26, 1986.

RISH, I. An empirical study of the naive Bayes classifier. **IJCAI 2001 workshop on empirical methods in artificial intelligence**, v. 3, p. 41–46, 2001.

SAFAVIAN, R.; LANDGREBE, D. A Survey of Decision Tree Classifier. **IEEE Transactions on Systems, Man, and Cybernetics**, v. 21, n. 3, p. 660–674, 1991.

SEBASTIANI, F. Machine Learning in Automated Text Categorization. **ACM Computing Surveys**, v. 34, 2002.

SILVA, E. L. DA; MENEZES, E. M. **Metodologia da Pesquisa e Elaboração de Dissertação**. 4. ed. Florianópolis: UFSC, 2005.

SINAPSE. **Sinapse da Inovação.** Disponível em: <http://sc.sinapsedainovacao.com.br/>. Acesso em: 5 de jun. 2019.

THORLEUCHTER, D.; VAN DEN POEL, D.; PRINZIE, A. Mining ideas from textual information. **Expert Systems with Applications**, v. 37, n. 10, p. 7182–7188, 2010.

VIJAYAN, V. K.; BINDU, K. R.; PARAMESWARAN, L. **A Comprehensive Study of Text Classification Algorithms**. p. 1109–1113, 2017.

WAZLAWICK, R. S. Uma Reflexão sobre a Pesquisa em Ciência da Computação à Luz da Classificação das Ciências e do Método Científico. **Revista de Sistemas de Informação da FSMA**, n. 6, p. 3–10, 2010.

ZHANG, H. **The Optimality of Naive Bayes**. AA 1.2, 2004.

ZHANG, H.; LI, D. **Naïve Bayes Text Classifier**. n. 3, p. 708–711, 2007.