UTILIZAÇÃO PRÁTICA DE WORD EMBEDDING APLICADA À CLASSIFICAÇÃO DE TEXTO

Luiz Fernando Spillere de Souza[1];
Alexandre Leopoldo Gonçalves[2]

***Abstract**: Text classification aims to extract knowledge from unstructured text patterns. The concept of word incorporation is a representation technique that allows words with similar meanings to have a similar representation, in order to incorporate reasoning characteristics about their use and meaning. The aim of this article is to analyze the work already published on the use of embedded words applied to the classification of texts, to propose a practical application that demonstrates its effectiveness. This study contributes to proving the effectiveness of the use of word incorporation applied to text classification, having reached an accuracy rate of around 73%.*
Keywords: Text Classification, Word Embeddings, Knowledge Extraction.

Resumo: A classificação de texto visa extrair conhecimento de padrões de texto não estruturados. O conceito de incorporação de palavras é uma técnica de representação que permite que palavras com significados semelhantes tenham uma representação semelhante, a fim de incorporar características de raciocínio sobre seu uso e significado. O objetivo deste artigo é a partir da análise dos trabalhos já publicados abordando o uso de palavras incorporadas aplicadas à classificação de textos, propor uma aplicação prática que demonstre sua eficácia. Este estudo contribui para a comprovação da eficácia da utilização da incorporação de palavras aplicada à classificação de texto tendo atingido um índice de acurácia em torno de 73%.
Palavras-chave: Classificação de Texto, Incorporação de Palavras, Extração de Conhecimento.

[1] Doutorando no Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento (PPGEGC) - Universidade Federal de Santa Catarina (UFSC) - Florianópolis - SC - Brasil. Email: spillere@gmail.com
[2] Professor Titular no Departamento de Engenharia e Gestão do Conhecimento (dEGC) – Universidade Federal de Santa Catarina (UFSC) - Florianópolis – SC – Brasil. Email: a.l.goncalves@ufsc.br

1 INTRODUÇÃO

A classificação de texto é uma tarefa de processamento e recuperação de informações justificada pela crescente disponibilidade de informações em formato digital que, por consequência, gera necessidade de acessar estas informações de maneira eficiente e ordenada.

O objetivo principal da classificação de texto é extrair conhecimento a partir de padrões do texto não estruturado de várias fontes. É uma área de pesquisa que assume o desafio de produzir ferramentas para a tomada de decisão, analisar grandes quantidades de texto em linguagem natural e encontrar padrões (Brindha, Prabha, & Sukumaran, 2016).

Para classificar adequadamente um texto não estruturado é necessária uma compreensão sintática e semânita do texto, além do conhecimento de uma quantidade suficientemente grande de categorias de classificação (Aliyeva, Kim, Choi, & Lee, 2018). Os métodos tradicionais de classificação de texto também necessitam que todo documento seja convertido em um vetor numérico. Especificamente, cada palavra (termo) que aparece no documento faz parte de um vetor que em sua totalidade representa um documento. (Pinheiro, Cavalcanti, & Ren, 2015).

Devido a grande quantidade de categorias necessárias e alta dimensão destes vetores de representação dos termos, os algoritmos de mineração de dados enfrentam problemas de ruído, dados esparsos e alta dimensionalidade (Hoai Nam & Quoc, 2017). A seleção de características de texto é uma maneira eficaz para reduzir a dimensão do modelo de espaço vetorial, onde o objetivo é selecionar um subconjunto reduzido, porém com forte capacidade de representação do conjunto de palavras originais (Zhu, Wang, & Zou, 2017)

A partir desta constatação, surge então o método *Word Embeddings* (WE), que consiste em uma abordagem capaz de identificar as informações semânticas latentes da linguagem, capturando a coocorrência de padrões das palavras (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Pennington, Socher, & Manning, 2014), o qual permite a redução da alta dimensionalidade nos dados e o raciocínio sobre o uso e significado de palavras (Shuang, Zhang, Loo, & Su, 2020).

O objetivo deste artigo consiste na análise dos trabalhos já publicados abordando o uso de palavras incorporadas aplicadas à classificação de textos e, a partir disso, propor uma

aplicação prática que demonstre sua eficácia. Este estudo contribui para a comprovação da eficácia da utilização da incorporação de palavras aplicada à tarefa de classificação de texto. A partir desta comprovação deseja-se futuramente incrementar o conhecimento e criar um método mais abrangente de classificação de texto.

2 ASPECTOS CONCEITUAIS

2.1 CLASSIFICAÇÃO DE TEXTO

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 atualmente.

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 Altınel e Ganiz (2018) definem a classificação automática de texto como a tarefa de organizar documentos em classes pré-determinadas, 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).

2.2 WORD EMBEDDINGS

O *Word Embedding* ou incorporação de palavras é uma forma de representação que permite que palavras com significados semelhantes também tenham uma representação semelhante. Isso possibilita que as máquinas desenvolvam uma melhor compreensão das palavras (Aubaid & Mishra, 2018). É uma abordagem poderosa para capturar as informações

semânticas latentes da linguagem, que podem ser utilizadas para classificação de texto, capturando a coocorrência de padrões das palavras, permitindo incorporar características de raciocínio sobre o uso e significado das palavras (Mikolov, 2013).

Uma das características do *Word Embedding* é a utilização de vetores densos e de baixa dimensão cujo benefício é o aumento da capacidade computacional, uma vez que não há necessidade de se manipular vetores dispersos e de alta dimensionalidade, como é o caso de alguns outros classificadores (Goldberg, 2013). As representações podem ser facilmente construídas a partir de vetores de incorporação de palavras que também têm a vantagem de não necessitarem de uma grande quantidade de documentos para treinar os modelos que promovem suporte à tarefa classificação de texto (Sinoara, Camacho-Collados, Rossi, Navigli, & Rezende, 2019).

Os *Word Embedding* são modelos matemáticos que codificam relações de palavras dentro de um espaço vetorial. Eles são criados através do processo de treinamento baseado em informações de coocorrência entre palavras. Constitui-se em uma ferramenta emergente para o processamento de linguagem natural sendo utilizada em uma ampla variedade de tarefas de processamento de idiomas. Sua utilidade decorre da capacidade de codificar relacionamentos de palavras no espaço vetorial. As aplicações variam desde componentes em sistemas de processamento de linguagem natural até ferramentas para análise linguística no estudo de linguagem e literatura (Heimerl & Gleicher, 2018).

2.3 USO DO WORD EMBEDDING NA CLASSIFICAÇÃO DE TEXTO

O *Word Embedding* pode ser usado como um método de pré-processamento na classificação de texto. Algumas propriedades justificam sua aplicação na classificação de texto, sendo as principais: refletem a similaridade e dissimilaridade entre as palavras (Mikolov, Chen, Corrado, & Dean, 2013) e incorporam as palavras de um texto e suas características em vetores de baixa dimensionalidade (Dhillon, Foster, & Ungar, 2015).

Uma das implementações mais atuais do *Word Embedding* é o *Word2vec*. O *Word2vec* é uma ferramenta baseada em *deep learning* (aprendizado profundo) criado por uma equipe de pesquisadores liderada por Tomas Mikolov (Mikolov, Chen, et al., 2013) e lançada pelo

Google® em 2013. O modelo *Word2Vec* usa as informações de contexto de uma palavra para convertê-la em um vetor de baixa dimensão, permitindo a capacidade de representar relacionamentos semânticos (Qiu et al., 2018).

2.4 TRABALHOS RECENTES QUE UTILIZAM WORD EMBEDDING NA CLASSIFICAÇÃO DE TEXTO

Na literatura existem vários estudos comprovando a aplicação bem sucedida do *Word Embedding* na classificação de textos, dentre os quais citaremos alguns trabalhos que evidenciam a eficácia desta utilização.

Um modelo de linguagem neural, baseado em tópicos *Skip-gram, Word Embedding* e em Redes Neurais Convolucionais foi proposto por (Xu et al., 2016) para classificar dados textuais biomédicos, capturando as relações semânticas das palavras com modelos de tópicos. As palavras indexadas no *Word Embedding* são utilizadas como entradas às arquiteturas de Redes Convolucionais Multimodais. Os experimentos conduzidos em vários conjuntos de dados do mundo real mostram que a combinação proposta executa com sucesso as tarefas de classificação de textos, incluindo a indexação de artigos médicos.

Outra abordagem para a classificação de texto com base em *Word Embedding* rurgiu inspirada em "*Bag of Visual Words*" que é um modelo de palavras, amplamente utilizado em visão computacional. Os autores Butnaru e Ionescu (2017) relatam ter obtidos bons resultados em duas tarefas de mineração de texto, a saber: categorização de texto por tópico e classificação de polaridade.

Na sua utilização considerando a semântica das palavras e incorporando a representação implícita do texto, o *Word Embedding* implementado no *Word2vec* foi utilizado na tarefa de classificação de texto baseada no *Open Directory Project*. Ao contrário do uso comum do *Word2vec*, foram utilizados os vetores de entrada e saída e isso permitiu calcular uma combinação típica de similaridade entre palavras, o que é mais eficaz na classificação do texto (Aliyeva et al., 2018).

A incorporação de palavras também atua em conjunto com a classificação de texto baseada em regras e o trabalho de Aubaid e Mishra (2018) concentrou-se principalmente nas

áreas sociais: ciências, classificação de produtos para compras, bibliotecas digitais e filtragem de *spam*. Os resultados contribuíram para determinar a boa atuação dos sistemas baseados em regras, bem como fornecer um guia para auxiliar os pesquisadores em planejamento de pesquisas futuras.

Em Kilimci e Akyokus (2018) foi utilizado diferentes representações de documentos com o *Word Embedding* e um conjunto de classificadores de base para classificação de texto. O conjunto de classificadores de base inclui algoritmos de aprendizado de máquina, como *Naive Bayes*, *Support Vector Machine*, *Random Forest* e uma *Deep Learning-Based Conventional Network*. Foi realizada a análise da precisão da classificação de diferentes representações de documentos empregando um conjunto de classificadores e os resultados experimentais demonstraram que o uso de métodos de aprendizado profundo em conjunto com *Word Embedding* melhoram o desempenho da classificação dos textos. O autor cita como classificadores de melhor desempenho o *Random Forest* e a *Deep Learning-Based Conventional Network*.

Um método geral de pré-processamento então foi proposto para cenários em que os dados de treinamento são escassos. Ele agrupa termos semanticamente semelhantes através do *Word Embedding*, que simulam como humanos pré-processam textos, substituindo palavras desconhecidas por termos conhecidos e também agrupando palavras semanticamente semelhantes (Elekes, Di Stefano, Schaler, Bohm, & Keller, 2019).

Neste outro trabalho foram realizados experimentos adicionando técnicas de *Word Embedding*, usando não só o *Word2vec* mas também testando o *Doc2vec*, para um conjunto de classificação de textos clínicos. Os resultados foram comparados com o uso do método tradicional *Bag Of Words*. O estudo mostrou que as técnicas de incorporação de palavras tiveram um desempenho melhor que o método *Bag Of Words* (Shao, Taylor, Marshall, Morioka, & Zeng-Treitler, 2019).

Um novo método de representação de texto denominado *Hybrid Word Embeddings* foi proposto por Song, Srimani e Wang (2019) combinando informações semânticas e contextuais obtidas do *Wordnet e* extraídas de documentos de texto, para fornecer representações concisas e precisas dos textos. O estudo experimental sobre classificação de

documentos mostra que o método proposto supera os métodos existentes, incluindo o *Doc2Vec* e *Word2Vec*, em termos de precisão de classificação.

3 PROCEDIMENTOS METODOLÓGICOS

O método científico utilizado segue o método indutivo que parte do princípio que o pesquisador define 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). Quanto à natureza é uma pesquisa aplicada, pois objetiva gerar conhecimentos para aplicação prática e dirigidos à solução de problemas específicos (Silva & Menezes, 2005). Os resultados da pesquisa poderão ser quantificados, recorrendo à linguagem matemática para descrever as causas do fenômeno e as relações entre variáveis, seguindo portanto uma abordagem quantitativa (Fonseca, 2002). 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 e em seu desenvolvimento 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) Limpar caracteres especiais e *Stop Words* da base de dados;
4) Extração das 5 palavras de maior frequência de ocorrência em cada texto da base de dados;
5) Submissão das palavras à um vocabulário previamente treinado através de *Word Embedding*, tendo como resultado a sua codificação equivalente;
6) Simplificação do grau da codificação das palavras obtidas para 3 com a finalidade de visualização posterior em 3 dimensões. Desta forma, cada texto da base de dados foi resumido em 5 pontos de 3 dimensões;
7) Encontrar um figura geométrica (bloco) mínima, formada pelos pontos de cada texto e seu baricentro;

8) Plotagem dos baricentros calculados, de forma que se possa separar por cores os textos contendo ideias e os textos comuns;
9) Verificação se existe proximidade entre os pontos.

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 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 aleatóriamente da Wikipedia®.

Para implementação dos *Word Embedings* foi utilizado o *Word2Vec*, que é um método de incorporação de palavras proposto por Mikolov et al. (2013). Tem como princípio a aprendizagem de vetores dimensionais usando um dos dois modelos neurais distintos: *Continuous Bag of Words (CBOW)* ou *Skip-Gram*. O *CBOW* prevê uma palavra atual com base em seu contexto, que corresponde às palavras vizinhas. Já o *Skip-Gram* busca a predição do contexto dado uma palavra. Neste trabalho optou-se pela utilização do *CBOW*, que de acordo com Mikolov et al. (2013) é mais rápido e funciona bem com palavras frequentes.

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Para ilustrar os procedimentos definidos na seção de procedimentos metodológicos, será utilizado um dos textos caracterizado como ideia para demonstrar toda a sua codificação.

O texto possui o seguinte conteúdo: "*Criar um APP voltado a gestão de pequenas propriedades rurais com alimentação de dados e controle de desempenho, nas dimensões financeira e produtiva (desempenho produtivo e zootécnico). O objetivo da ideia é facilitar a gestão em propriedades rurais, com ênfase em duas dimensões principais: 1) desempenho econômico: rentabilidade, lucratividade, margem de contribuição, ponto de equilíbrio, indicadores, etc. e 2) desempenho produtivo: identificação e acompanhamento produtivo nas*

diversas atividades de uma pequena propriedade rural, acompanhando aspectos de produtividade e zootécnicos (controle de animais vacinas peso etc.). A proposta visa oferecer alternativa de gestão por intermédio de um aplicativo que facilite o controle contínuo de dados financeiros e produtivos zootécnicos, considerando custos despesas e receitas da propriedade rural em suas diferentes atividades, gerando informações para a tomada de decisões. Embora haja muita tecnologia e recursos disponíveis no mercado, as pequenas propriedades rurais ainda não dispõem de muitos recursos para o controle de suas atividades, fragilizando a tomada de decisões e em consequência os resultados de suas atividades, uma vez que na grande maioria dos casos não possuem informações minimamente sistematizadas. Dessa perspectiva, resultam também limitações na qualidade de vida das famílias rurais. O desafio pensado segue na linha de uma inovação incremental, onde se pretende superar a falta de alternativas tecnológicas simplificadas de apoio a gestão de pequenas propriedades rurais. O oferecimento da solução de interesse de grande número de pessoas, famílias rurais, as quais em geral já dispõem de smartfones, o que facilitaria o uso de um aplicativo voltado a controlar o desempenho das atividades das propriedades. A contribuição esperada é de importante impacto social e poderá oportuniza maior produtividade e melhor qualidade de vida no campo."

As palavras-chave e sua codificação *Word Embedding* de ordem 3, conforme estabelecido na etapa 6 dos procedimentos metodológicos, estão na tabela 1:

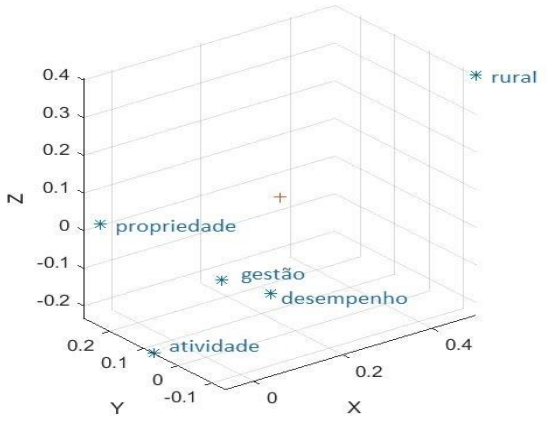
Tabela 1- Codificação *Word Embedding* do texto exemplo contendo uma ideia.

rural	propriedade	desempenho	atividade	gestão
0,500257	-0,027241999	0,068833999	-0,06357	0,196664006
-0,13873	0,275985003	-0,09973	0,07132	0,210511997
0,406158	0,004379	-0,040605001	-0,23274	-0,193737999

Fonte: autor

A partir disso, é traçada então uma figura geométrica (bloco) de forma que consiga englobar de forma mínima os 5 pontos das coordenadas das palavras. Encontra-se então o baricentro desta figura, conforme estabelecido na etapa 7 dos procedimentos metodológicos, demonstrado na figura 1.

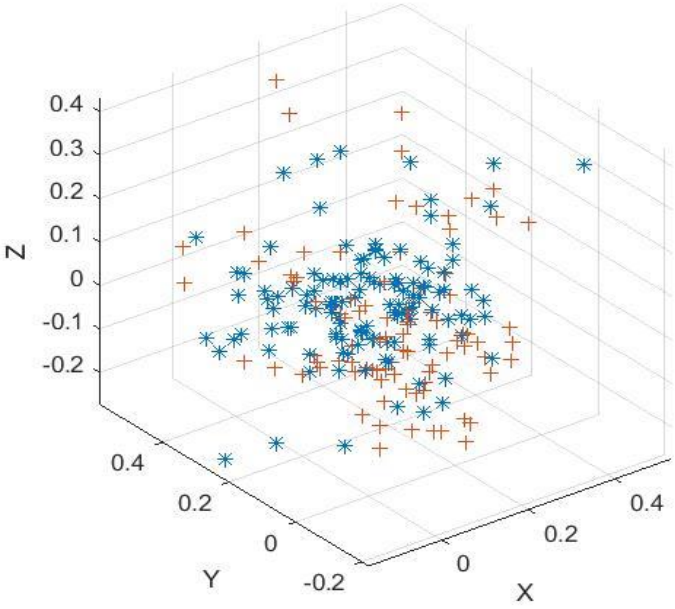
Figura 1 – Coordenadas das palavras-chave do exemplo de ideia e seu baricentro



Fonte: autor

Este procedimento é realizado para cada um dos 122 textos contendo ideias e também 100 textos contendo textos comuns, conforme estabelecido na etapa 8 dos procedimentos metodológicos. O resultado é plotado na figura 2, de forma que para sua diferenciação, os textos comuns são mostradas no símbolo '+' em vermelho e os textos com ideias são mostrados no símbolo '*' em azul.

Figura 2 – Plotagem das ideias e textos comuns



Fonte: autor

A partir deste gráfico pode-se verificar que há uma maior proximidade entre os pontos correspondentes aos textos que contem ideias, demonstrando ser possível uma diferenciação matemática dos textos. Para comprovar tal verificação visual foi encontrado o ponto médio dos baricentros das ideias e calculadas as médias das distâncias entre o baricentro de cada ideia versus seu ponto médio, onde o resultado das médias foi de 0,092. O mesmo foi feito para os textos comuns e seu resultado foi 0,109. Esta diferenciação entre as médias mostra que os baricentros dos textos contendo ideias apresentam uma menor distância entre si, mostrando-se mais densos.

Foi calculado também o baricentro geral dos textos das ideias e traçou-se uma figura geométrica delimitadora que pudesse otimizar esta separação entre os textos contendo ideias e textos comuns. Utilizando o baricentro geral dos textos das ideias como ponto central, a figura foi traçada variando suas distancias delimitadoras dos eixos X, Y e Z. Os valores foram obtidos após testes com diferentes configurações, demonstrados na tabela 2.

Tabela 2- Difrentes configurações da figura delimitador e acurácia obtida.

Identificador do Teste	Distância somada ao baricentro geral dos textos das ideias no eixo X	Distância somada ao baricentro geral dos textos das ideias no eixo Y	Distância somada ao baricentro geral dos textos das ideias no eixo Z	Acurácia
1	0,1	0,1	0,1	59%
2	0,2	0,2	0,2	68%
3	0,3	0,3	0,3	65%
4	0,4	0,4	0,4	56%
5	0,2	0,3	0,1	71%
6	0,3	0,2	0,1	69%
7	0,25	0,26	0,12	73%

Fonte: autor

A melhor configuração obtida no teste identificado como 7 possibilitou a obtenção da matriz de confusão demonstrada na tabela 3.

Tabela 3- Matriz de Confusão entre ideias e textos comuns.

	ideia	texto comum
ideia	86 (71%)	36 (29%)
texto comum	26 (26%)	74 (74%)

Fonte: autor

A matriz de confusão mostra que o índice de acerto dos textos contendo ideias foi de 71% e dos textos comuns de 74%, atingindo uma acurácia total em torno de 73%.

5 CONSIDERAÇÕES FINAIS

A partir da análise prática demonstrada neste estudo, pode-se verificar que o método de incorporação de palavras (WE) se mostra eficaz na tarefa de classificação de texto, promovendo a separação entre os textos contendo ideias dos textos comuns. Isso é esperado, pois o propósito dos WE são aglomerar palavras que contenham significados semelhantes, logo textos semelhantes também devem ter a tendência de ficarem aglomerados.

Este estudo serve de base para um estudo futuro que objetiva criar um método mais abrangente de classificação de texto e mineração de ideias. A classificação de texto promovida pelo WE servirá de base para definir graus de pertinência de um texto, tendo como base a distância entre o texto analisado e a aglomeração dos textos contendo ideias.

6 REFERÊNCIAS

Aliyeva, D., Kim, K. M., Choi, B. J., & Lee, S. K. (2018). Combining Dual Word Embeddings with Open Directory Project Based Text Classification. *Proceedings of 2018 IEEE 17th International Conference on Cognitive Informatics and Cognitive Computing, ICCI*CC 2018*, 179–186. https://doi.org/10.1109/ICCI-CC.2018.8482044

Altınel, B., & Ganiz, M. C. (2018). Semantic text classification: A survey of past and recent advances. *Information Processing and Management*, *54*(6), 1129–1153. https://doi.org/10.1016/j.ipm.2018.08.001

Aubaid, A. M., & Mishra, A. (2018). Text classification using word embedding in Rule-based methodologies: A systematic mapping. *TEM Journal*, *7*(4), 902–914. https://doi.org/10.18421/TEM74-31

Brindha, S., Prabha, K., & Sukumaran, S. (2016). A survey on classification techniques for text mining. *ICACCS 2016 - 3rd International Conference on Advanced Computing and Communication Systems: Bringing to the Table, Futuristic Technologies from Arround the Globe*, *01*(i), 1–5. https://doi.org/10.1109/ICACCS.2016.7586371

Butnaru, A. M., & Ionescu, R. T. (2017). From Image to Text Classification: A Novel

Approach based on Clustering Word Embeddings. *Procedia Computer Science*, *112*, 1783–1792. https://doi.org/10.1016/j.procs.2017.08.211

Dhillon, P. S., Foster, D. P., & Ungar, L. H. (2015). Eigenwords: Spectral word embeddings. *Journal of Machine Learning Research*, *16*, 3035–3078.

Elekes, A., Di Stefano, A. S., Schaler, M., Bohm, K., & Keller, M. (2019). *Learning from Few Samples: Lexical Substitution with Word Embeddings for Short Text Classification*. 111–119. https://doi.org/10.1109/jcdl.2019.00025

Fonseca, J. J. S. da. (2002). *Metodologia da pesquisa científica* (pp. 1–127). pp. 1–127. https://doi.org/10.1063/1.4740259

Goldberg, Y. (2013). Neural Network Methods for Natural Language Processing. *Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International*, *37*, 0–2. https://doi.org/10.1162/COLI

Heimerl, F., & Gleicher, M. (2018). Interactive Analysis of Word Vector Embeddings. *Computer Graphics Forum*, *37*(3), 253–265. https://doi.org/10.1111/cgf.13417

Hoai Nam, L. N., & Quoc, H. B. (2017). Integrating Low-rank Approximation and Word Embedding for Feature Transformation in the High-dimensional Text Classification. *Procedia Computer Science*, *112*, 437–446. https://doi.org/10.1016/j.procs.2017.08.058

Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*, *52*(1), 273–292. https://doi.org/10.1007/s10462-018-09677-1

Kilimci, Z. H., & Akyokus, S. (2018). Deep learning- and word embedding-based heterogeneous classifier ensembles for text classification. *Complexity*, *2018*. https://doi.org/10.1155/2018/7130146

Lakatos, E. M., & Marconi, M. de A. (2010). *Metodologia científica* (2nd ed.; Atlas, Ed.). São Paulo: Atlas.

Mikolov, T. (2013). Learning Representations of Text using Neural Networks. *NIPS Deep Learning Workshop*, 1–31.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. 1–12. Retrieved from http://arxiv.org/abs/1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed

representations ofwords and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 1–9.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Pinheiro, R. H. W., Cavalcanti, G. D. C., & Ren, T. I. (2015). Data-driven global-ranking local feature selection methods for text categorization. *Expert Systems with Applications*, *42*(4), 1941–1949. https://doi.org/10.1016/j.eswa.2014.10.011

Qiu, J., Chai, Y., Liu, Y., Gu, Z., Li, S., & Tian, Z. (2018). Automatic Non-Taxonomic Relation Extraction from Big Data in Smart City. *IEEE Access*, *6*, 74854–74864. https://doi.org/10.1109/ACCESS.2018.2881422

Shao, Y., Taylor, S., Marshall, N., Morioka, C., & Zeng-Treitler, Q. (2019). Clinical Text Classification with Word Embedding Features vs. Bag-of-Words Features. *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, 2874–2878. https://doi.org/10.1109/BigData.2018.8622345

Shuang, K., Zhang, Z., Loo, J., & Su, S. (2020). Convolution–deconvolution word embedding: An end-to-end multi-prototype fusion embedding method for natural language processing. *Information Fusion*, *53*(May 2019), 112–122. https://doi.org/10.1016/j.inffus.2019.06.009

Silva, E. L. da, & Menezes, E. M. (2005). *Metodologia da Pesquisa e Elaboração de Dissertação* (4th ed.). Florianópolis: UFSC.

Sinapse. (2017). Sinapse da Inovação. Retrieved June 5, 2019, from http://sc.sinapsedainovacao.com.br/

Sinoara, R. A., Camacho-Collados, J., Rossi, R. G., Navigli, R., & Rezende, S. O. (2019). Knowledge-enhanced document embeddings for text classification. *Knowledge-Based Systems*, *163*, 955–971. https://doi.org/10.1016/j.knosys.2018.10.026

Song, X., Srimani, P. K., & Wang, J. Z. (2019). Hwe: Hybrid word embeddings for text classification. *ACM International Conference Proceeding Series*, 25–29. https://doi.org/10.1145/3342827.3342837

Vijayan, V. K., Bindu, K. R., & Parameswaran, L. (2017). *A Comprehensive Study of Text*

Classification Algorithms. 1109–1113.

Wazlawick, R. S. (2010). 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*, *nº 6*, 3–10.

Xu, H., Kotov, A., Dong, M., Carcone, A. I., Zhu, D., & Naar-King, S. (2016). Text classification with topic-based word embedding and Convolutional Neural Networks. *ACM-BCB 2016 - 7th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, 88–97. https://doi.org/10.1145/2975167.2975176

Zhu, L., Wang, G., & Zou, X. (2017). Improved information gain feature selection method for Chinese text classification based on word embedding. *ACM International Conference Proceeding Series*, 72–76. https://doi.org/10.1145/3056662.3056671