Comparação das Técnicas de Extração de Características em Texto Aplicado à Mineração de Opinião.

Marilla Eduarda Leite Nepomuceno[1], Pedro Henrique Melchiori Tridico[2],
Orientador: Prof. Dr. Henrique Dezani[3]

e-mail:
mleite.nepo@gmail.com[1]; pedrohenrimt@gmail.com[2];
henrique.dezani@fatec.sp.gov.br[3]

Resumo: Este artigo consiste na comparação das extrações de características na mineração de opinião utilizando um *corpus* de avaliações online de clientes. A comparação aconteceu entre os métodos *Bag of Words*, TF-IDF e *Word2Vec*, os resultados do tempo de execução e precisão do modelo apresentados através da matriz de confusão levaram à conclusão que a melhor técnica varia de acordo com o objetivo da análise.

Palavras-chave: Aprendizado de Máquina, Extração de Característica, Mineração de Opinião, Processamento de Linguagem Natural.

Abstract: *This article consists of comparing the extractions of characteristics in opinion mining using a corpus of online customer reviews. The comparison occurred between the Methods Bag of Words, TF-IDF and Word2Vec, the results of the execution time and accuracy of the model presented through the confusion matrix led to the conclusion that the best technique varies according to the purpose of the analysis.*

Keywords: *Feature Engineer, Machine Learning, Natural Language Processing, Opinion Mining.*

1. Introdução

A expansão da internet foi um fenômeno que possibilitou a popularização do *e-commerce* (comércio eletrônico) na economia mundial. O *e-commerce* é caracterizado por qualquer transação comercial realizada por plataformas digitais, desde a escolha do produto ou serviço, até o pagamento e finalização do pedido (VÁZQUEZ, 2021).

Neste tipo de comércio, encontramos também o *marketplace*, cada vez mais consolidado. Segundo Alvarenga (2021) trata-se de um espaço online para divulgação de inúmeros tipos de produtos em suas plataformas, em troca de um percentual sobre as vendas. Como grandes exemplos de *marketplace* atuantes no Brasil temos o Mercado Livre, Americanas.com, Magazine Luiza e Olist, na qual o último se faz o objeto de estudo deste artigo.

A Olist é uma das maiores lojas de departamento atuante no *e-commerce* brasileiro, e tem como foco lojistas das mais diversas categorias de produtos. Surgiu em 2007 chamando-se Solidarium, com objetivo de divulgar o trabalho de artesãos e artesãs de Curitiba-PR. A loja sempre buscou parceria com varejistas como Walmart, e desde 2011 passa a ser um *marketplace*. Em 2015 passou a ser Olist, com uma inovação, uma loja de departamentos nos *marketplace* (OLIST, 2021).

O crescimento das plataformas gera banco de dados robustos, os quais ficam suscetíveis a ruídos que reduzem a qualidade, e consequentemente a eficiência dos resultados após um processo de mineração de dados (SILVA, 2014). Porém, segundo Vinodhini e Chandrasekaran (2014), as avaliações de produtos interpretam papel fundamental nas análises de mercado das empresas, uma vez que opiniões de consumidores impactam diretamente na venda do produto.

Portanto, para que essa análise seja realizada é preciso dispor de ferramentas como aprendizado de máquina e métodos de processamento de linguagem natural, como mineração de opinião para processar massivas quantidades de dados.

Na mineração de opinião existem diferentes métodos de extração de características com foco na classificação, frequência e contexto. Apesar de possuírem o mesmo objetivo, as abordagens podem não representar o mesmo resultado através de variações na precisão e qualidade dos dados. Neste contexto foi realizada uma comparação dos métodos utilizando o mesmo *corpus* a fim de determinar a melhor abordagem para análise de sentimentos.

2. Objetivo Geral

Realizar a mineração de opiniões, comparando as técnicas de extração de características usadas na área de processamento de linguagem natural e aprendizado de máquina a partir de um conjunto de dados público da Olist disponível no Kaggle.

3. Objetivos Específicos

- Realizar a extração de características utilizando *Bag of Words*;
- Realizar a extração de características utilizando *TF-IDF*;
- Realizar a extração de características utilizando *Word2Vec* nas arquiteturas *Skip-Gram* e *Continuous Bag of Words*;
- Realizar treinamento de um modelo em regressão logística para minerar opinião usando as características extraídas nas etapas anteriores;
- Apresentar e discutir os resultados obtidos em cada técnica utilizada.

4. Fundamentação Teórica

Neste capítulo será abordada a fundamentação teórica referente aos conceitos de mineração de opinião, aprendizado de máquina, extração de característica e processamento de linguagem natural, como *Bag of Words*, TF-IDF, *Word2Vec* (*Skip-Gram* e CBOW).

4.1 Mineração de Opinião

A mineração de opinião, também conhecida como análise de sentimentos, análise de subjetividade e análise de opinião, é a área de estudo que realiza análises acerca de avaliações, opiniões ou sentimentos de clientes sobre determinado serviço, empresas ou produtos. As emoções compartilhadas podem ser negativas ou positivas (LIU, 2012).

De acordo com o mesmo autor os estudos surgidos sobre a mineração de opinião, têm grande ligação com o crescimento das redes sociais. Os comentários disponibilizados nos mesmos, geram cada vez mais conteúdos que são valiosos tanto para clientes, mas também em maior importância para empresas, e suas tomadas de decisões.

4.2 Aprendizado de Máquina

Conforme Monard e Baranauskas (2003), o aprendizado de máquina é uma área da inteligência artificial, que possui como objetivo a elaboração de sistemas e técnicas capazes de adquirir conhecimento e tomar decisões baseadas nas experiências adquiridas durante problemas anteriores. Dentro da área do aprendizado de máquina (AM), temos abordagens de aprendizado supervisionado e não supervisionado.

A abordagem supervisionada, utiliza de uma técnica de apresentação de diversos exemplos, induzindo assim o modelo a ser capaz de classificar novas instâncias, com base no

modelo de treino. Já a abordagem não supervisionada, consiste na utilização de dados não categorizados. Essa abordagem é indicada quando não é necessário um modelo preditivo, mas um modelo para encontrar dados regulares (CARVALHO, 2014).

4.3 Extração de Característica

Para que os objetivos propostos na mineração de opinião sejam alcançados é necessária a realização de algumas tarefas específicas. Dentro delas temos a extração de características. Essa tarefa, corriqueiramente, é dividida em outras duas, devido às características diferentes presentes em cada atributo. Portanto, podemos dividir em Extração de Entidades e Extração de Aspectos (VIEIRA, 2018).

As entidades, geralmente, são encontradas em nome de produtos, indivíduos e empresas. Já os aspectos, são os atributos referenciados das entidades mencionadas.

Por exemplo no relato “Excelente mochila, entrega super-rápida” as palavras “excelente” e “rápida” indicam um sentimento acerca da entidade “mochila”. Com isso, verificamos que a extração das características auxilia nas etapas de análise para obtenção de resultados significativos.

4.4 Processamento de Linguagem Natural

Entramos agora em um conceito muito importante para o desenvolvimento do estudo em questão, afinal o que é o Processamento de Linguagem Natural (NPL)?

Segundo Thanaki (2017), o NPL é um ramo da Inteligência Artificial, e trata-se da capacidade das tecnologias computacionais e linguísticas em realizar o processamento da linguagem natural humana. Alguns serviços já em utilização pelo público, em que se encontra presente o PLN, são por exemplo o Google Assistente do Google e a Siri Speech da Apple.

Alguns ramos que utilizam dos conceitos de processamento de linguagem natural são: aprendizado de máquina citado acima; programação; análise de *corpus* que é um conjunto de documentos sobre um determinado tema.

4.4.1 *Bag of Words*

O *Bag of Words* (bolsa de palavras) é uma das técnicas utilizadas na PLN. Neste modelo, os dados contidos no *corpus* estão na forma de texto, e são apresentados em “*bags*” de palavras, não considerando nesse caso a gramática ou ordem das palavras (THANAKI, 2017).

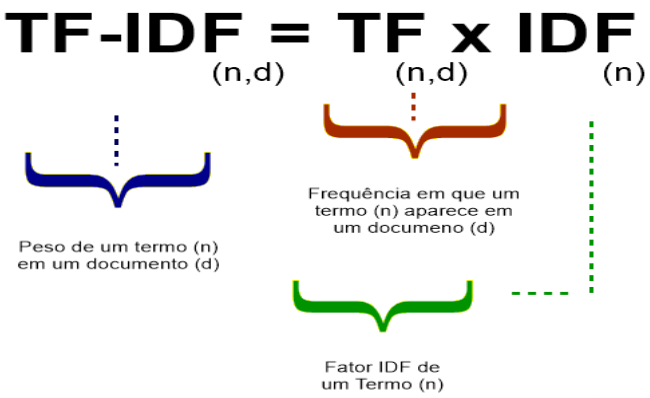
É um modelo que pode ser utilizado para o pré-processamento dos textos, que quando armazenada nas "*bags*", faz uma contagem do total de ocorrências das palavras usadas com mais frequência.

4.4.2 TF-IDF

TF-IDF (*Term Frequency-Inverse Document Frequency*) é um modelo simples, porém de grande utilidade. Em linhas gerais, através dele é indicado quantas vezes determinado termo (palavra) aparece em um conjunto de dados e qual sua importância para análise do *corpus* e definição dos resultados (THANAKI, 2017).

Para entender melhor, vamos analisar a Figura 1 e as fórmulas presentes nesse método. O termo TF indica a frequência de cada palavra (TF t) = (Número de vezes que o termo t aparece no documento) / (Número total de termos no documento)). Por outro lado, a sequência IDF indica realmente a importância das palavras, palavras mais comuns tendem a ser consideradas como menos significativas. Para o cálculo do termo, utiliza-se o logaritmo de base 2: IDF(t) = log (N/df(t)).

Figura 1 Exemplificação da fórmula TF-IDF



Fonte: AMD3 Marketing

4.4.3 *Word2Vec*

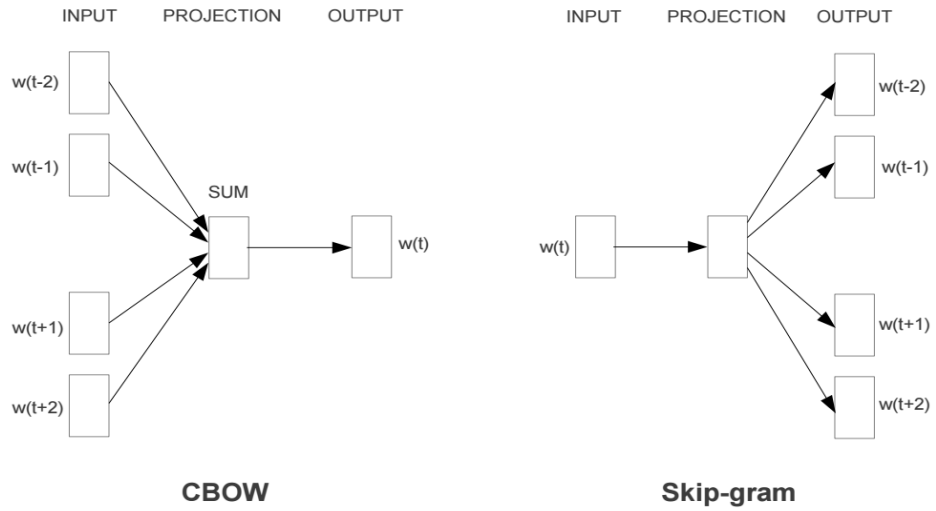
O termo *Word2Vec*, tem como objetivo a procura por representações vetoriais através do uso de redes neurais. A finalidade tem como ideia analisar o poder preditivo das palavras próximas (CAVALCANTI et al., 2017).

Consistem em modelos que geram a incorporação de palavras em vetores. Os modelos possuem camadas de redes neurais sendo uma camada de entrada, uma camada escondida e

uma camada de saída. Existem duas arquiteturas do *Word2Vec: Skip-Gram* onde o dado de entrada é a palavra central e a saída as palavras de contexto; e o *Continuous Bag of Words* (CBOW) onde os dados de entrada são as palavras de contexto e a saída a palavra central das frases. A esquematização das arquiteturas pode ser conferida na Figura 2.

Quando comparado ao modelo *Bag of Words*, o *Word2Vec* possui uma maior capacidade preditiva maior, devido ao uso das redes neurais (MASSONI, 2021).

Figura 2: Comparação das arquiteturas Word2Vec



Fonte: HackDeploy

5. Trabalhos Similares

Estudo comparativo de Mineração de Opiniões em rede varejista

O artigo compara os resultados obtidos por uma análise realizada a partir de técnicas de mineração de textos em conteúdos extraídos de uma rede de comércio varejista com os resultados obtidos por uma pesquisa de campo feita para a mesma rede varejista. O trabalho corrobora com a ideia de que as técnicas de Mineração de Textos, em particular a Análise de Sentimentos, apresentam-se cada vez mais como soluções eficientes e vantajosas para conhecimento de opiniões e estudos de percepção e reputação de marcas.

Hecksher e Ebecken (2016) de forma similar a este artigo utilizam a mineração de opinião para extrair as características do *corpus* de rede varejista.

Mineração de opinião usando princípio de análise de componentes principais baseada em análise modelo de conjunto para aplicação de e-commerce

O artigo explora e compara a aplicação híbrida do aprendizado de máquina e técnicas de redução para classificação de opinião. Vinodhini e Chandrasekaran (2014) utilizaram métodos similares de mineração de opinião baseados em estatística como a regressão logística para comparar a qualidade de classificação das opiniões em positivas e negativas das avaliações de produtos de e-commerce.

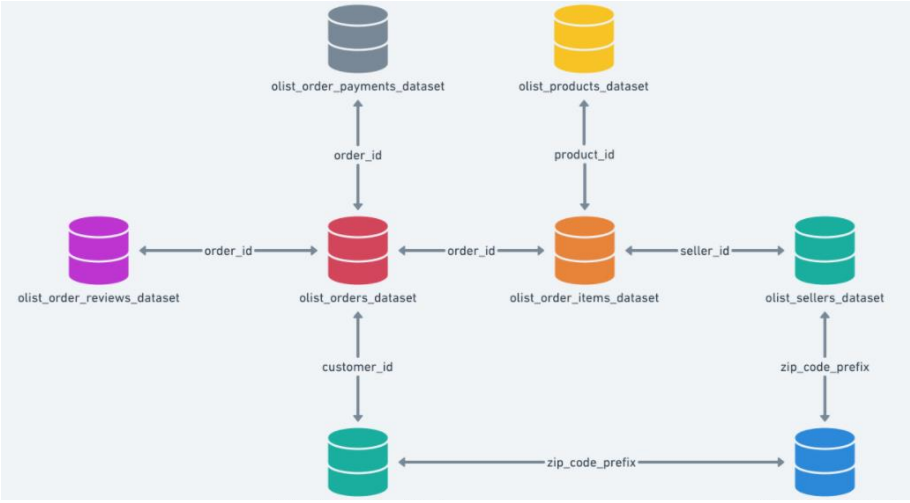
6. Metodologia

6.1 População, coleta e amostra de dados

Os dados foram retirados do dataset público do E-commerce Olist disponibilizado pela plataforma Kaggle.

O dataset possui 100.000 (cem mil) registros entre 2016 e 2018. Para o artigo foram utilizadas apenas as informações das avaliações dos produtos contidas no olist_order_reviews_dataset. A Figura 3 apresenta a esquematização do banco.

Figura 3 Esquematização do banco de dados



Fonte: Kaggle

6.2 Métodos a serem utilizados

Com o *corpus* pré-processado foram desenvolvidos quatro algoritmos para as técnicas de extração de características baseadas em estatísticas e redes neurais: *Bag of Words*, *TF-IDF*,

Word2Vec (arquiteturas *Skip-Gram e CBOW*). Cada característica foi treinada utilizando a regressão logística e o resultado apresentado através da matriz de confusão.

6.3 Ferramentas e tecnologias utilizadas

Para realização do artigo foi utilizada a plataforma Google Colab para programação na Linguagem Python. As bibliotecas utilizadas foram Pandas para manipulação dos dados; *Gensim* e *Scikit-Learn* para criação da extração e característica e treinamento do modelo; *Matplotlib* para visualização dos dados.

7. Desenvolvimento

7.1 Pré Processamento

O *corpus* foi fornecido pela plataforma Kaggle, para este projeto utilizou-se apenas o arquivo composto por avaliações de clientes em compras realizadas na Olist apresentado na Figura 4.

Figura 4: Demonstração da tabela utilizada

Fonte: Kaggle

7.1.1 Alterações no Corpus

Com finalidade de apresentar uma análise com o máximo de precisão e devido ao mesmo conter uma grande quantidade de dados, realizou-se o pré-processamento dos dados presentes no dataset. Dentre as estratégias de limpeza dos dados manteve-se apenas as colunas *"review"* e *"review_comment_message"* apresentadas na Figura 5.

Além disso ocorreu a retirada de registros que possuíam valores nulos em qualquer uma das colunas citadas e ao fim a conversão da escala da coluna "review" de 0-5 para 0-1, sendo valor igual ou acima de 3 considerado uma avaliação positiva, portanto 1 e abaixo de 3 avaliação negativa, ou seja, 0.

Figura 5 Apresentação dos registros da tabela comentários

review_id	order_id	review	review_comment_title	review_comment_message	review_creation_date	review_answer_timestamp
e64fb393e7b328	658677c97b3	5		Recebi bem antes do prazo es	2017-04-21 00:00:00	2017-04-21 22:02:06
f7c4243c7fe1938	8e6bfb81e283	5		Parabéns lojas lannister adorei	2018-03-01 00:00:00	2018-03-02 10:26:53
8670d52e15e00(	b9bf720beb4a	4	recomendo	aparelho eficiente. no site a ma	2018-05-22 00:00:00	2018-05-23 16:45:47
4b49719c8a200(	9d6f15f95d01e	4		Mas um pouco ,travando...pelo	2018-02-16 00:00:00	2018-02-20 10:52:22
3948b09f7c818e2	e51478e7e27	5	Super recomendo	Vendedor confiável, produto ok	2018-05-23 00:00:00	2018-05-24 03:00:01
9314d6f9799f5bf	0dacf04c5ad5	2		GOSTARIA DE SABER O QUE	2018-01-18 00:00:00	2018-01-20 21:25:45

Fonte: Kaggle

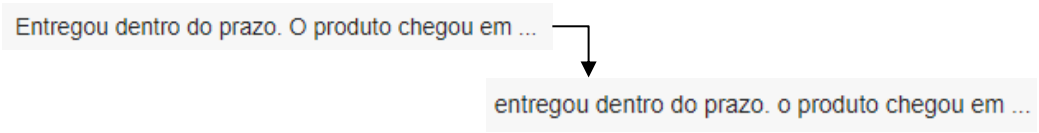
Após todas as mudanças realizadas os dados da coluna *"review"* foram balanceados para permanecer com a mesma quantidade de valores positivos e negativos, finalizando o pré-processamento de dados com aproximadamente 22 mil registros.

7.1.2 Tratamentos de Texto

Quando uma grande variedade de informações é analisada, além do processamento de limpeza também é necessário tratar para que o texto seja o mais padrão e legível ao computador. Por isso, foram realizadas tratativas de normalização e padronização do texto presentes nos comentários de clientes.

A Figura 6 apresenta a primeira técnica utilizada, cujo objetivo é transformar todo o texto em caixa baixa (*lowercase)*, de forma que todo o texto esteja em letras minúsculas.

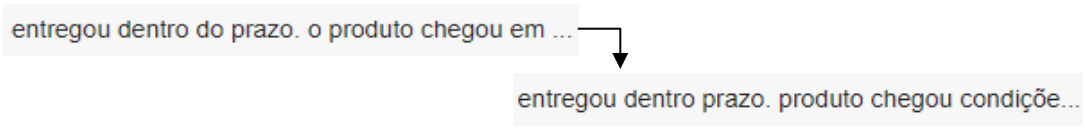
Figura 6 Demonstração técnica *lowercase*



Fonte: Os autores

Seguindo a Figura 7 demonstra que foi retirado as palavras vazias (*stopwords)* como preposições, que não possuem valor real para a análise.

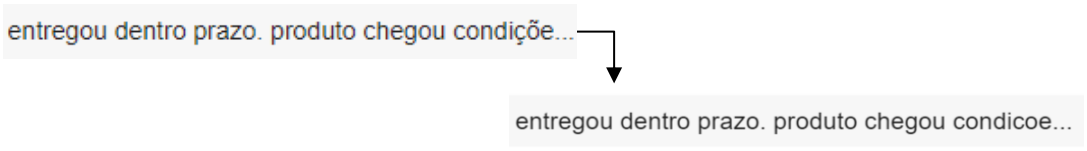
Figura 7 Demonstração técnica *stopwords*



Fonte: Os autores

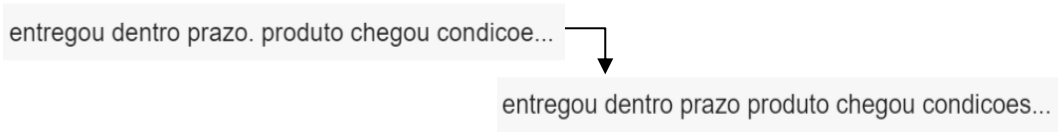
Na Figura 8 apresenta a retirada dos acentos (*unidecode*) e Figura 9 da pontuação (*punctuation*)

Figura 8 Demonstração técnica *unidecode*



Fonte: Os autores

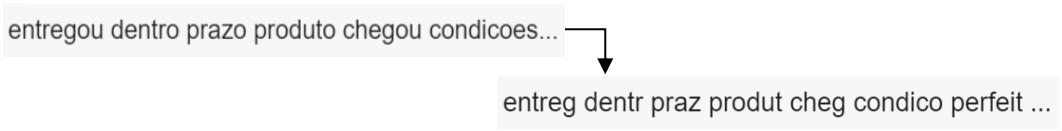
Figura 9 Demonstração técnica *punctuation*



Fonte: Os autores

Por fim, a rotina do *stemming* é demonstrada na Figura 10 reduzindo as palavras nas suas raízes, como por exemplo, “entregou” para “entreg”, buscando o máximo de similaridade e padronização dos textos.

Figura 10 Demonstração técnica *stemming*



Fonte: Os autores

7.2 Treinamento Regressão Logística

Segundo Gonzalez (2018) a regressão logística é uma técnica para determinar, a partir de um conjunto de variáveis independentes contínuas e/ou binárias uma saída categórica binária (sim/não).

Esse modelo pode ser usado para prever a probabilidade desse evento ocorrer de acordo com as informações analisadas. Na regressão logística é comumente utilizada uma variável que com o nome de Y que carrega com o resultado do treinamento.

A técnica estatística foi utilizada para treinamento do modelo devido a sua eficiência e versatilidade de não depender de grande número de recursos computacionais para predição probabilística.

O projeto treinou os modelos na seguinte ordem:

- A variável X carrega o resultado do processo de extração de característica.
- Foi determinado que da quantidade de dados existentes em X, 20% seriam destinados ao teste final do modelo e os 80% o treinamento. A separação dos dados entre teste e treino foi determinada de forma randômica pelo sistema.
- Realizou o treinamento na técnica de regressão logística

7.3 Métrica

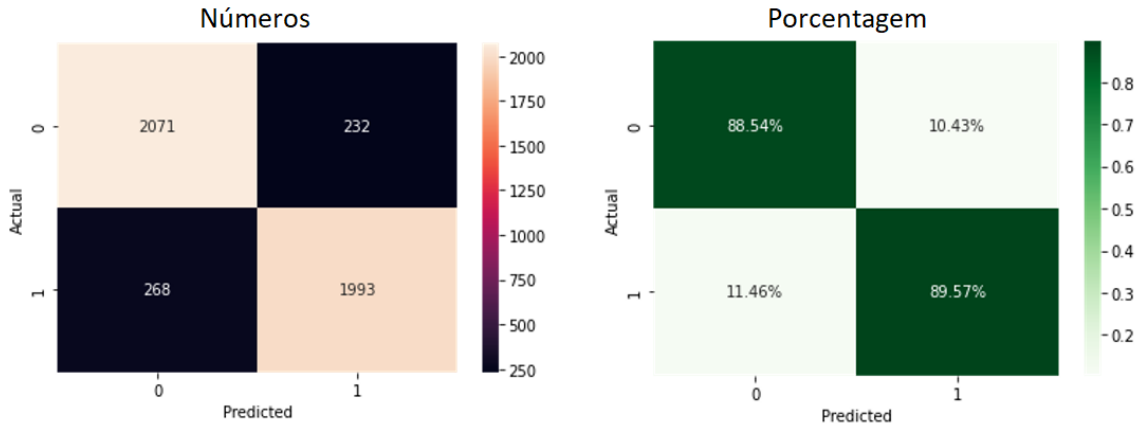
Para obtenção dos resultados e análise dos modelos, utilizou-se do método de Matriz de Confusão. Segundo Prina e Trentin (2015), a matriz de confusão é uma forma de representação através da relação entre os dados de referência (dados verdadeiros e de treinamento) com dados classificados. Através do mesmo podemos obter quatro resultados, *"True Positive"*, onde o modelo prediz uma informação verdadeira corretamente, *"False Negative"*, por meio da qual o modelo prediz uma informação falsa incorretamente. Temos ainda o *"False Positive",* predizendo uma informação falsa corretamente, e por fim, o *"True Negative"*, gerando informações trazidas como verdadeiras, porém incorretamente.

8. Resultado e Discussão

Entre os resultados que podemos evidenciar através desse trabalho, destacamos o modelo de classificação obtido através da matriz de confusão, acima mencionada. Sendo assim, obteve-se os seguintes dados:

8.1 *Bag of Words*:

Figura 11 Matriz de Confusão referente ao Bag of Words em valores reais e porcentagem

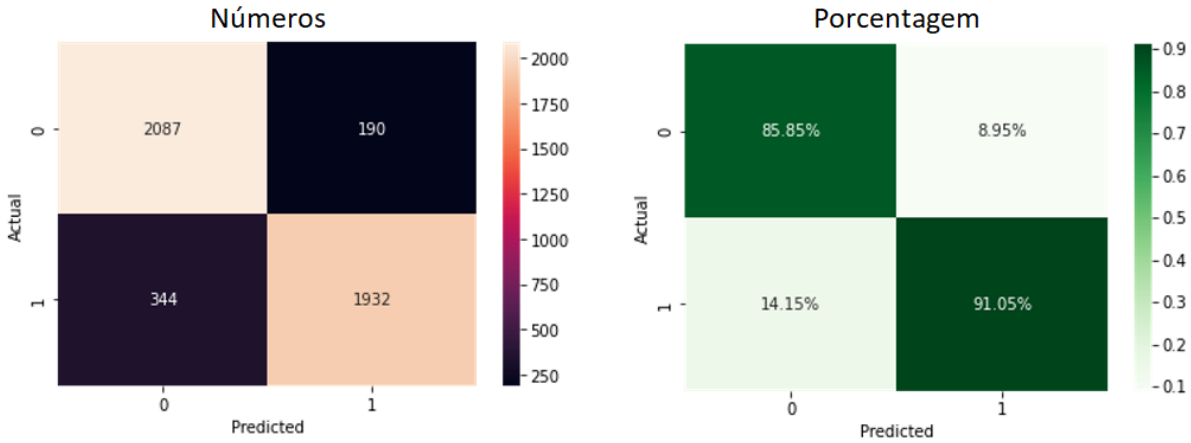


Fonte: Os autores

O método do *Bag of Words* levou um total de 225 milissegundos para criação com uma base de aproximadamente 22 mil registros. Durante a criação do modelo de aprendizado de máquina 80% desses registros foram utilizados como treino, sendo os 20% restantes o teste final para medir a precisão do modelo. O tempo para criação, treinamento e teste desse modelo foi de 395 milissegundos. O resultado apresentou uma taxa de acerto de 88,54% para as avaliações negativas e 89,57% de avaliações positivas. Portanto considera-se o modelo final com a precisão e acurácia de 89%.

8.2 TF-IDF:

Figura 12 Matriz de Confusão referente ao TF IDF em valores reais e porcentagem



Fonte: Os autores

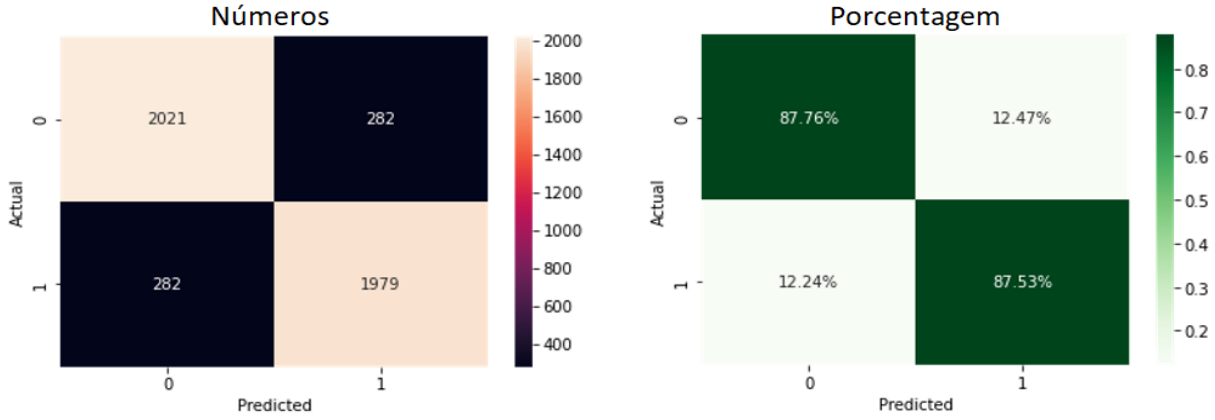
O método do TF-IDF possibilitou a visualização de palavras não existentes no vocabulário, por exemplo onomatopeias de “aaaaa”, “xxxx” e caracteres especiais, para

melhorar a qualidade dos dados foi realizada uma limpeza dessas palavras. O método levou um total de 2.56 segundos para criação com uma base de aproximadamente 22 mil registros. A criação do modelo de aprendizado de máquina seguiu o padrão 80% treino e 20% teste final de precisão do modelo. O tempo de treinamento e teste desse modelo foi de 229 milissegundos. O resultado apresentou uma taxa de acerto de 85,85% para as avaliações negativas e 91,05% de avaliações positivas. Portanto considera-se o modelo final com a precisão de 91% e acurácia de 88%. Ambos os métodos são considerados mais simples e pode se perceber um resultado semelhante e naturalmente alto de ambos, a principal divergência notada do TF-IDF foi a facilidade em notar palavras e caracteres que atrapalham o modelo não tratados anteriormente.

8.3 *Word2Vec*:

ScripGram:

Figura 13 Matriz de Confusão referente ao Word2Vec – Skip-Gram em valores reais e porcentagem

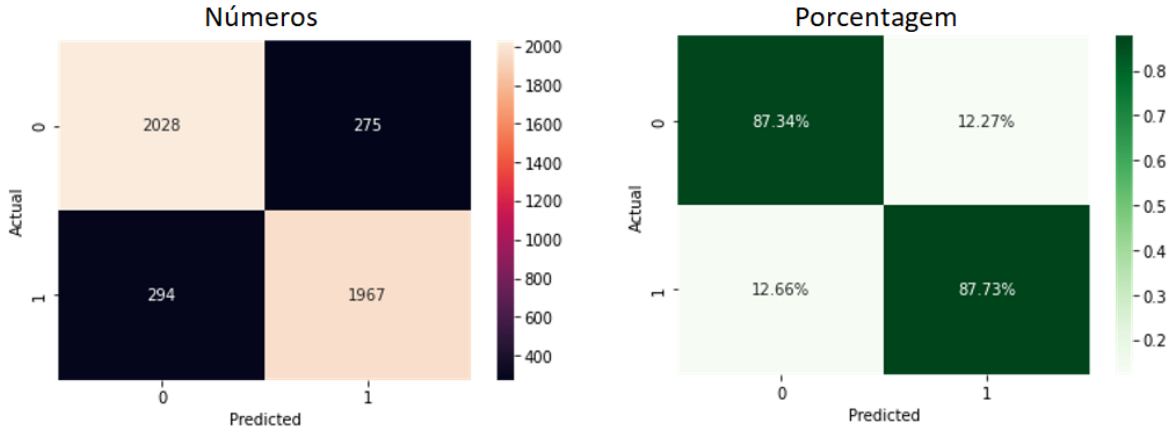


Fonte: Os autores

O método do *Word2Vec* utilizando a arquitetura *Skip-Gram* em que a entrada é a palavra central do texto (palavra que carrega o contexto da frase) seguiu as seguintes formatações: considerou-se como intervalo de palavras contexto o valor 2 (que significa duas palavras antes e duas palavras depois da palavra central) e o vetor de palavras sendo trezentos. O método demorou 35.5 segundos para criação com uma base de aproximadamente 22 mil registros. A criação do modelo de aprendizado de máquina seguiu o padrão 80% treino e 20% teste final de precisão do modelo. O tempo para treinamento e teste desse modelo foi de 1.56 segundos. O resultado apresentou uma taxa de acerto de 87,76% para as avaliações negativas e 87,53% de avaliações positivas. Portanto considera-se o modelo final com a precisão e acurácia de 88%.

CBOW:

Figura 14 Matriz de Confusão referente ao Word2Vec - CBOW em valores reais e porcentagem



Fonte: Os autores

O método do *Word2Vec* utilizando a arquitetura CBOW em que a entrada são as palavras de contexto seguiu as seguintes formatações: considerou-se como intervalo de palavras contexto o mesmo que o método *Skip-Gram*, o valor 2 (duas palavras antes e duas palavras depois da palavra central) e o vetor de palavras sendo trezentos. O método demorou 19.3 segundos para criação com uma base de aproximadamente 22 mil registros. A criação do modelo de aprendizado de máquina seguiu o padrão 80% treino e 20% teste final de precisão do modelo. O tempo para treinamento e teste desse modelo foi de 1.22 segundos. O resultado apresentou uma taxa de acerto de 87,34% para as avaliações negativas e 87,73% de avaliações positivas. Portanto considera-se o modelo final com a precisão e acurácia de 88%. Ambos os modelos de *Word2Vec* possuem formatação e resultados semelhantes, sendo a maior divergência o tempo do método sendo que o *Skip-Gram* demorou aproximadamente 50% a mais que o CBOW para executar.

8.4 Resumo

Tabela 1 Comparação de precisão e tempo dos métodos

		Tempo de execução em milissegundos		
Extração de característica	**Precisão**	**Extração de característica**	**Treino modelo**	**Total**
Bag of Words	89%	225	395	**620**
TF-IDF	**91%**	2.540	229	2.769
Word2Vec Skip-Gram	88%	35.500	1.560	37.060
Word2Vec CBOW	88%	19.300	1.220	20.520

Fonte: Os autores

Ambas as arquiteturas do *Word2Vec* apresentaram precisão semelhante e ligeiramente abaixo dos resultados encontrados nos métodos anteriores do TF-IDF e *Bag of Words*, o que pode levar a crer que apesar de complexo o Word2Vec não é possui maior vantagem em relação aos anteriores. De fato, se o objetivo da análise é apenas ser capaz de diferenciar comentários positivos e negativos as técnicas mais simples suprem essa necessidade. Porém se o intuito for uma abordagem contextual do texto o método mais complexo é o que possui as ferramentas capazes.

9. Conclusão

A expansão do comércio eletrônico possibilitou o surgimento de novas oportunidades de mercado como por exemplo, a modalidade do *marketplace*. Em grandes empresas dessa modalidade são gerados dados robustos a respeito da opinião do cliente de onde é possível extrair grande valor, desde que sejam aplicadas técnicas com intuito de filtrar e extrair as características como opinião e perfil de compra contidas nesses bancos de dados.

As técnicas de análise de sentimento utilizadas no artigo são o primeiro passo em direção ao entendimento da visão do cliente em relação aos produtos e em consequência a própria empresa.

Existem vários caminhos a serem seguidos com o mesmo objetivo, porém é preciso que seja claro quais os resultados esperados para encontrar o método que se encaixe no perfil.

De forma geral, verifica-se que todas as técnicas de extração de dados tiveram um desempenho satisfatório com precisões muito semelhantes.

Levando em consideração o *trade off* do tempo de execução a técnica do *Bag of Words* foi 78% mais rápida que a TF-IDF e 98% mais rápida que as técnicas do Word2Vec, porém é um método limitado para visualizar e manipular os dados. O TF-IDF apresentou a melhor precisão de 91%, pode ser considerado um meio termo entre as opções citadas por ser tão simples quanto o *Bag of Words* de ser executado e com possibilidades de visualização de dados mais abrangentes. Tratando-se do objetivo de realizar análises contextuais a técnica que possui melhor capacidade analítica é o *Word2Vec* que apresenta uma ínfima diferença de precisão, entre 1% e 3%, além de trabalhar no campo de redes neurais.

Por isso, é perceptível que cada extração apesar de serem utilizadas com o mesmo objetivo possuem características distintas, sendo necessário assim ter definido o objetivo e resultados esperados desses modelos além de levar em consideração quantidade, qualidade e linguagem dos dados a serem analisados, para assim escolher assertivamente o modelo que cumpre os requisitos.

10. Referências Bibliográficas

ALVARENGA, Marcos Vinícius de. **Visualização de Dados e sua Importância para a Tomada de Decisão: Uma Aplicação Usando o Conjunto de Dados de E-Commerce da Olist**. 2021. 43 f. Monografia (Especialização) - Curso de Estatística, Departamento de Estatística, Universidade Federal de Ouro Preto, Ouro Preto, 2021. Disponível em: https://monografias.ufop.br/bitstream/35400000/2994/6/MONOGRAFIA_Visualiza%c3%a7%c3%a3oDadosImport%c3%a2ncia.pdf. Acesso em: 30 ago. 2021.

AM3D MARKETING. **Entenda o Uso da Densidade de Palavras-Chave no SEO Aplicando a TF-IDF**. 2018. Disponível em: https://am3dmarketing.wordpress.com/2018/08/05/entenda-o-uso-da-densidade-de-palavras-chave-no-seo-aplicando-a-tf-idf%E2%80%8B%E2%80%8B%E2%80%8B%E2%80%8B%E2%80%8B%E2%80%8B%E2%80%8B/. Acesso em: 04 nov. 2021.

CARVALHO, Hialo Muniz. **Aprendizado de Máquina voltado para Mineração de Dados: Árvores de Decisão**. 2014. 100 f. Monografia (Especialização) - Curso de Engenharia de Software, Universidade de Brasília, Brasília, 2014. Disponível em: https://bdm.unb.br/bitstream/10483/9487/1/2014_HialoMunizCarvalho.pdf. Acesso em: 08 set. 2021

CAVALCANTI, Anderson et al. **Uma Nova Abordagem para Detecção de Plágio em Ambientes Educacionais.** In: XXVIII Simpósio Brasileiro De Informática Na Educação - Sbie (Brazilian Symposium On Computers In Education), 28., 2017, [S.L.]. Anais do XXVIII Simpósio Brasileiro de Informática na Educação (SBIE 2017). [S.L.]: Brazilian Computer Society (Sociedade Brasileira de Computação - Sbc), 2017. p. 1177-1186. Disponível em: http://www.br-ie.org/pub/index.php/sbie/article/view/7646/5442. Acesso em: 11 set. 2021.

GONZALEZ, Leandro de Azevedo. **Regressão Logística e suas Aplicações**. 2018. 45 f. Monografia (Especialização) - Curso de Ciência da Computação, Centro de Exatas e Tecnológicas, Universidade Federal do Maranhão, São Luís, 2018. Disponível em:

https://monografias.ufma.br/jspui/bitstream/123456789/3572/1/LEANDRO-GONZALEZ.pdf. Acesso em: 12 nov. 2021.

HACKDEPLOY. **Word2Vec Explained Easily**. 2018. Disponível em: https://www.hackdeploy.com/word2vec-explained-easily/. Acesso em 09 nov. 2021.

HECKSHER, Andrea; EBECKEN, Nelson F. F. Estudo Comparativo De Mineração De Opiniões Em Rede Varejista. **Revista Eletrônica Sistemas & Gestão**, Rio de Janeiro, v. 11, n. 4, p. 423-430, 2016. Disponível em: https://revistasg.uff.br/sg/article/view/1093/527. Acesso em: 14 jul. 2021.

LIU, Bing. **Sentiment Analysis and Opinion Mining**. [S. L.]: Morgan & Claypool Publishers, 2012. 180 p. Disponível em: https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf. Acesso em: 08 set. 2021.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. **Conceitos Sobre Aprendizado de Máquina. Sistemas Inteligentes Fundamentos e Aplicações**. 1 ed. Barueri-SP: Manole Ltda, 2003. p. 89--114. ISBN 85-204-168. Disponível em: https://dcm.ffclrp.usp.br/~augusto/publications/2003-sistemas-inteligentes-cap4.pdf. Acesso em: 08 set 2021.

OLIST (Curitiba). **DNA de tecnologia com coração de varejo**. 2021. Disponível em: https://olist.com/sobre-o-olist/. Acesso em: 30 ago. 2021.

OLIST; SIONEK, André. **Brazilian E-Commerce Public Dataset by Olist.** Kaggle. 2018. Doi: 10.34740/KAGGLE/DSV/195341. Disponível em: Brazilian E-Commerce Public Dataset by Olist | Kaggle. Acesso em: 30 ago. 2021.

PRINA, Bruno Zucuni; TRENTIN, Romario. 2015, João Pessoa. GMC: **Geração de Matriz de Confusão a partir de uma classificação digital de imagem do ArcGIS.** Santa Maria: Simpósio Brasileiro de Sensoriamento Remoto, 2015. Disponível em: http://www.dsr.inpe.br/sbsr2015/files/p0031.pdf. Acesso em: 20 nov. 2021.

SILVA, Michel de Almeida. **O Pré-Processamento em Mineração de Dados como método de suporte à modelagem algorítmica**. 2014. 83 f. Dissertação (Mestrado) - Curso de Modelagem Computacional de Sistemas, Universidade Federal do Tocantins, [Tocantins], 2014. Disponível em: http://download.uft.edu.br/?d=42f33aee-e6e1-4ff7-b93e-8266fa41782e;1.0:O%20Pr%C3%A9-Processamento%20em%20Minera%C3%A7%C3%A3o%20de%20Dados%20como%20m%C3%A9todo%20de%20suporte%20%C3%A0%20modelagem%20algor%C3%ADtmica.%20Dissert.%20SILVA,%202014.pdf. Acesso em: 30 ago. 2021.

VIEIRA, João Paulo Albuquerque. **Análise de Métodos de Extração de Aspectos em Opiniões Regulares.** 2018. 77 f. Dissertação (Mestrado) - Curso de Pós-Graduação em Ciência da Computação, Universidade Federal Do Piauí, Teresina, 2018. Disponível em: https://repositorio.ufpi.br/xmlui/bitstream/handle/123456789/1701/dissertacao%2bass.pdf?sequence=1. Acesso em: 08 set. 2021.

VINODHINI, G; CHANDRASEKARAN, R M. **Opinion mining using principal component analysis-based ensemble model for e-commerce application**. [S. L.]: CSI Publications, 2014. 11 p. Disponível em: https://link.springer.com/content/pdf/10.1007/s40012-014-0055-3.pdf. Acesso em: 1 set. 2021

VÁZQUEZ, Alejandro. **O que é e-commerce e como funciona?** 2021. Disponível em: https://www.nuvemshop.com.br/blog/o-que-e-ecommerce/. Acesso em: 30 ago. 2021.