Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Análise das Correspondências e da Obra Artística de Vincent Van Gogh Utilizando Mineração de Textos e Processamento de Imagens

Matheus Veleci dos Santos

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Orientador
Prof. Dr. Alexandre Zaghetto

Coorientador
Prof. Dr. Marcus Vinícius Chaffim Costa

Brasília
2019

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Análise das Correspondências e da Obra Artística de Vincent Van Gogh Utilizando Mineração de Textos e Processamento de Imagens

Matheus Veleci dos Santos

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Prof. Dr. Marcus Vinícius Chaffim Costa (Orientador)
FGA/UnB

Prof. Dr. Alexandre Ricardo Soares Romariz
ENE/UNB

Prof. Dr. Vinicius Ruela Pereira Borges
CIC/UNB

Prof. Dr. José Edil Guimarães
Coordenador do Curso de Engenharia da Computação

Brasília, 10 de Julho de 2019

Dedicatória

Este trabalho é dedicado as pessoas que estiveram ao meu lado ao longo de toda vida: meus pais João Carmelino dos Santos Filho e Iris Veleci da Silva Santos. Também dedico ao meu irmão Thiago Veleci dos Santos e minha querida namorada Luisa Arcoverde Bezerra Soares que não mediram esforços para me dar apoio e carinho.

Agradecimentos

Agradeço aos meu pais, **João Carmelino dos Santos Filho** e **Iris Veleci da Silva Santos**, por seu amor e apoio em tempo integral durante a caminhada da graduação.

Agradeço também, ao meu irmão **Thiago Veleci dos Santos** que não faltou com a sua genialidade em diversas conversas e momentos ao longo da vida. Também à minha namorada **Luisa Arcoverde Bezerra Soares** por tornar a vida mais simples e hamorniosa.

À todos os meus socios na Módulo12: **Maximillian Fan Xavier**, **Rafael Dias da Costa** e **Murilo Cerqueira Medeiros**. Agradeço ainda, minha família Italiana e meus amigos de fora do Brasil que sempre torceram por mim como se estivessem presentes fisicamente.

Por fim agradeço, todo o corpo docente dos departamentos de Engenharia Elétrica e Ciência da Computação inclusive o professor **Dr. Alexandre Zaghetto** e a professora **Dra. Mylene Queiroz Farias** que influenciaram meu percuso acadêmico pelo seu exmplo e trabalho.

Resumo

O pintor holandês Vincent Van Gogh é uma das personalidades artísticas mais importantes do século XIX. Sua vida conturbada e sua saúde mental instável durante a carreira artística nos leva a desenvolver ferramentas capazes de melhorar a curadoria sobre sua obra e vida.

Com o avanço dos algoritmos de aprendizado de máquina, neste trabalho, se propõe uma ferramenta de análise para vida do artista Vincent Van Gogh utilizando processamento de linguagem natural, assim como o processamento de imagens para relacionar sua fonte biográfica com suas pinturas.

Com a análise de sentimentos e emoções sobre as mais de 800 cartas e a extração de características de suas pinturas culminando em um nível de complexidade, esse trabalho apresenta uma similaridade entre os mesmos de forma a auxiliar no entendimento do processo criativo que marca a trajetória desse artista.

Palavras-chave: Processamento de Linguagem Natural, Van Gogh, Aprendizado de Máquina, Arte, Processamento de Imagens

Abstract

The Dutch painter Vincent Van Gogh is one of the most important artistic personalities of the 19th century. His troubled life and unstable mental health during his artistic career lead us to develop tools capable of improving curation over his work and life.

With the advancement of the machine learning algorithms, in this work, an analysis tool for the life of the artist Vincent Van Gogh is proposed using natural language processing, as well as the processing of images to relate his biographical source with his paintings.

With the analysis of feelings and emotions about the more than 800 letters and the extraction of characteristics of his paintings culminating in a level of complexity, this work presents a similarity between them in order to help in the understanding of the creative process that marks the trajectory of this artist.

Keywords: Natural Processing Language, Van Gogh, Machine Learning, Art, Image Processing

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

IA Inteligência Artificial.

NPL Processamento de Linguagem Natural.

Capítulo 1

Introdução

A arte é a principal maneira na qual o ser humano expressa suas emoções e sua história, sendo essa de forma escrita, visual ou teatral. Durante a construção das primeiras sociedades complexas, a arte foi o único meio de comunicação onde eram registrados momentos históricos e artefatos sobre as estruturas sociais de uma época em que se ausentava a utilização da tecnologia de informação presente atualmente. O entendimento da arte nos permite compreender melhor a história da humanidade, assim como os pensamentos e valores que grandes personalidades artísticas deixaram na história, impulsionando o desenvolvimento humano e científico ao longo dos anos.

Com o avanço da tecnologia e com a maior disponibilidade pública das obras de arte, conseguimos desenvolver ferramentas que ajudam curadores e historiadores a aprimorar seus estudos, que são, muitas vezes, baseados em conhecimentos manuais oriundos de séculos passados. Para esse trabalho, escolheu-se de uma grande personalidade do século XIX, Vincent Van Gogh, como fonte de um estudo exploratório aplicando mineração de dados.

Van Gogh é considerado um dos maiores pintores da história da arte ocidental, e, ainda atualmente, sua obra continua a desafiar a compreensão de muitos especialistas, devido as suas características marcantes como, por exemplo, as famosas pinceladas grossas de seus quadros, a sua visão única para a representação de uma cena e seus famosos autorretratos.

Devido à falta de reconhecimento durante os anos de sua carreira, tudo que sabemos sobre o pintor é baseado em uma única referência biográfica: o conjunto das mais de 800 cartas trocadas com seu irmão Theo. Outro fato que se sabe da vida de Van Gogh está na sua grande impopularidade e desvalorização por parte da família, além de ter sido reprovado e em peso pela elite artística europeia de sua época. Devido a esses fatores, o pintor só foi reconhecido como um grande personagem na história da humanidade após a sua morte, em 1890.

Um dos pontos mais interessantes sobre a trajetória de Van Gogh é a variação de

sua saúde mental. A falta de um diagnóstico psiquiátrico preciso, não disponível naquele tempo, fez com que o artista ficasse conhecido como louco, enfraquecendo sua opinião e obra. Portanto, o entendimento de passagens da vida do artista sempre foram alvos de pesquisas em diversas áreas acadêmicas com o intuito de entender melhor o seu perfil psicológico e como o mesmo pode ter influenciado em sua obra artística.

Para contribuir com a análise de sua vida, este trabalho tem como objetivo explorar os dados contidos na fonte biográfica única, buscando identificar os sentimentos e emoções envolvidas e compará-los com a complexidade de sua obra visual. Tendo isso em vista, aplicamos métodos de processamento de linguagem natural e processamento de imagens digitais, com o propósito de correlacionar as diversas características relacionadas aos seus quadros à sua personalidade.

Por ser um conjunto de dados nunca abordado da maneira apresentada por este trabalho, foi aplicado um *pipeline* de processamento, desde a sua aquisição e consolidação, até o uso de algoritmos de aprendizado de máquina, para a identificação de padrões sobre a vida de Van Gogh.

Para a melhor compreensão do trabalho, no Capítulo 2 descrevemos uma breve introdução sobre o conceito da arte e sua relação com as emoções e sentimentos humanos. Posteriormente, é exposto um resumo da biografia do artista para colaborar com a compreensão do trabalho.

Já no Capítulo 3, apresentamos os conceitos teóricos referentes as tecnologias empregadas ao longo desse trabalho, tais como o processamento de linguagem natural, o processamento de imagens digitais, o *pipeline* utilizado em aprendizado de máquina e os algoritmos de classificação não supervisionados. Seguidos dos trabalhos correlatos no Capítulo 4, que embasaram esse trabalho.

Em seguida, no Capítulo 5, é definida a metodologia proposta no trabalho, passando pelas fases de concepção, aquisição dos dados, pré-processamento, organização, busca de padrões e métricas utilizadas.

Os resultados obtidos com a aplicação da metodologia, prevista acima, podem ser visualizados no Capítulo 6, em conjunto com as respectivas análises. Por fim, o Capítulo 7, resulta de uma breve descrição do trabalho e define as possíveis atividades a serem realizadas no futuro.

Capítulo 2

Vincent Van Gogh

O objetivo desse capítulo é apresentar uma breve introdução sobre os conceitos de arte e como ela se relaciona com o ser humano. Em um segundo momento, é introduzida uma curta biografia do artista Vincent Van Gogh.

2.1 A Arte e suas Relações

O mundo moderno considera a arte como um das obras mais importantes da história da humanidade. A prova desse argumento é a quantidade de visitas em museus e exposições espalhados ao redor do mundo. Apesar disso, a busca pela definição de "arte"e a sua função continua nos levando a diversas teorias e contradições [6].

A arte consegue ampliar nossas capacidades para além dos limites originalmente impostos pela natureza e também, pelo momento histórico. Logo, em [6] define-se como um meio terapêutico que pode ajudar a guiar, incentivar e consolar tanto quem produz como quem é espectador. Portanto, esse instrumento presente ao longo da história da humanidade pode ser considerado como uma compensação de nossas fraquezas inatas. Neste caso, fraquezas mais mentais do que físicas, que podem ser chamadas de fragilidades psicológicas.

Além do seu conceito abstrato, podemos definir algumas funções da arte: a rememoração, a esperança, o sofrimento, o reequilíbrio, a compreensão de si, o crescimento e a apreciação.

A rememoração diz respeito a memória de um fato, seja ele histórico ou não. Antes da fotografia, o único meio de gravar ou registrar momentos icônicos era por meio da arte, mesmo após o advento de novas tecnologias.

A esperança está inserida na arte por meio da beleza representada a partir do nosso mundo. Apreciamos coisas belas e graciosas que alimentam a sentimentalidade que é o oposto da complexidade utilizada para designar algo problemático. Por outro lado, o

sofrimento é um aspecto que a arte nos ajuda a enfrentar como forma de expressão da dor de maneira digna.

Além das características já citadas, a arte também pode reestruturar nosso equilíbrio emocional, com doses concentradas de emoções tendo como uma de suas funções o reequilíbrio. Sua função como a real compreensão de si, visa um ponto de reflexão e de como nos enxergamos e reagimos à arte, seja observando obras belas ou obras mais realistas.

Às duas últimas funções estabelecidas por [6] definem nossa relação com o mundo e a sociedade. Quando falamos que sua função é o crescimento, isso vem de uma relação que temos com a arte, a mesma provoca diversas emoções como medo, tédio, felicidade e calma, nos levando a criar associações com pensamentos atuais de nossa sociedade. A outra principal função da arte é a apreciação, que diz respeito a perceber o que está ao nosso redor. A arte remete a comparações com o mundo real e imaginário o que nos leva a conclusão sobre o nosso mundo.

Todas as funções e seus conceitos mostram o quão importante é a arte em nossa sociedade e como compreendê-la é de extrema importância. Logo, quando conseguimos entender a arte, estamos percebendo o ser humano a partir de suas relações psicológicas, seja com ele mesmo, com o mundo ou com a sociedade.

2.2 Vincent Van Gogh

O pintor holandês Vincent Van Gogh é uma das personalidades artísticas mais importantes do século XIX, entretanto, seus quadros atemporais com características únicas e marcantes foram objetos de valor apenas após sua morte. A curta carreira artística, de apenas 9 anos, revela uma vida difícil e uma personalidade com diversos distúrbios psicológicos que influenciaram seu não reconhecimento pela crítica da arte europeia da época. Suas obras são estudadas para tentarmos entender o momento histórico e o processo criativo de Van Gogh, considerando sua história pessoal. Nesse tópico, será apresentada uma pequena biografia para entendermos mais sobre sua história, todos os trechos descritos a seguir foram retiradas de [7] e [1].

Vincent Van Gogh, pintor holândes, nasceu no dia 30 de março de 1853, na cidade de Zundert, localizada no sul da Holanda. Sua mãe, Anna Carbentus e seu pai Theodorus, pastor protestante, tiveram seis filhos. Van Gogh era o filho mais velho e seus irmãos se chamavam: Anna, Theo, Wil, Lies e Cor. Aos treze anos, Vincent Van Gogh foi para uma escola secundária em Tilburg. Nesta, obteve excelentes notas, especialmente na área de humanas. Apesar disso, ele deixou a escola por razões desconhecidas e nunca mais voltou aos estudos. Quando completou dezesseis anos, seu tio, cujo nome também era Vincent, conseguiu um estágio para Van Gogh como revendedor de arte na filial da

Galeria Internacional de Arte Goupil & Cie, em Haia. Analisando a vida de Van Gogh, pode-se perceber que as primeiras cartas trocadas entre ele e seu irmão mais novo, Theo, são datadas deste período na Goupil & Cie. Posteriormente, Theo também começou a trabalhar na mesma galeria, porém, na cidade de Bruxelas.

Vincent Van Gogh teve uma breve passagem em Londres, conhecendo diversos museus renomados como: *Britsh Museum* e *National Gallery*. Neste último, o pintor teve contato com obras de François Millet e Jules Breton. Em 1875, quando foi transferido para Paris, o artista se tornou extremamente religioso. Por conta dessa influência religiosa, as cartas desse período são marcadas por citações bíblicas e sermões da igreja, além de referências a Deus.

Apesar de seu interesse pela arte, Van Gogh estava cada vez menos animado pelo seu trabalho como vendedor na galeria, tendo inclusive diversos problemas com clientes e funcionários da loja. Dessa forma, a Goupil acabou dispensando seus serviços. Após esse episódio, em 1877, seu tio Vincent conseguiu lhe um emprego em uma livraria, em Dordrecht. Entretanto, seus pais estavam muito preocupados com seu filho mais velho, visto que, até o momento, Van Gogh não havia demonstrado um propósito profissional e pessoal. Sendo assim, acabaram decidindo apoiá-lo na carreira religiosa, tendo Van Gogh estudado para o teste de admissão na área teológica.

(a) (b)

Figura 2.1: (a) uma gravura do artista Van Gogh na sua infânica e em (b) o primeiro desenho feito pelo artista [1].

Assim como anteriormente, mesmo com todo esforço, Vincent Van Gogh não tinha disciplina para os estudos, abandonando-os novamente. No entanto, ainda continuou sua devoção, tendo experiências de missões em pequenos vilarejos. Em 1880, incentivado pelo seu irmão Theo, Van Gogh decidiu se concentrar seus estudos na vida artística. A inspiração veio de seus desenhos enviados nas cartas que o artista trocou com seu

irmão. Por conta disso, mudou-se para Bruxelas, onde deu início aos estudos na técnica de desenhos, além de ter contato com outros artistas. Van Gogh não passou muito tempo morando em Bruxelas, voltando a morar com os pais no ano seguinte. A continuidade e prática com seus desenhos se deu, principalmente, ao ar livre. Esse fato é visto como uma marca do artista. Nesta época, como o pintor não possuía trabalho remunerado, Theo, gerente da Goupil & Cie Paris na época, enviava dinheiro para o sustentar o irmão mais velho.

Todo esse cenário, bem como a escolha da carreira eram motivos de preocupação dos pais de Vincent Van Gogh, pois, para eles a vida de artista era vista socialmente como um fracasso, ao contrário de uma profissão honrada. Esse assunto explicava o fato da relação entre pais e filho ser conturbada, já que existia uma reprovação dos pais pelas suas escolhas pessoais.

Outro ponto marcante na vida de Vincent Van Gogh são as aulas de pintura com seu primo, o artista Anton Mauve. Tendo como intuito melhorar suas técnicas com a pintura, Van Gogh passou a praticar fanaticamente. Além disso, com o auxílio de seu primo, desenvolveu novas habilidades de perspectiva e pintura em aquarela e óleo. Com a desenvoltura de suas pinturas, Van Gogh enviou alguns exemplares para que seu irmão tentasse vender em Paris. Contudo, apesar da expectativa, as vendas não foram bem sucedidas, dado que o gosto francês da época remetia a uma arte mais viva, isto é, com a presença de cores, diferentemente da arte de Van Gogh, que era representada por tons mais escuros e com muitas sombras.

Nesse contexto, a convivência entre o pintor e seus pais se tornou cada vez pior, devido ao seu comportamento considerado inadequado e não convencional. Em 1885, com a morte de seu pai, Vincent mudou-se para um estúdio, onde começou a trabalhar na famosa pintura 'The Potato Eaters'. Naquele mesmo ano, Van Gogh decidiu matricular-se na academia de arte em Antuérpia, deixando a Holanda para nunca mais voltar. Como era uma academia tradicional, o artista não se adequou e logo mudou-se para Paris, onde seu irmão mais novo morava e trabalhava. Por conta da influência da época e de seu contato, por exemplo, com a arte de Claude Monet e Emile Bernard, o artista começou experimentar novos tons e estilos, dando espaço para cores vibrantes em suas obras e suas famosas pinceladas curtas e grossas. Além disso, seu trabalho também sofreu grande influência das xilogravuras japonesas, vendidas em Paris, neste período.

Depois de alguns anos em Paris, Vincent começou a rejeitar o estilo de vida de cidade grande, desejando a paz do campo, o sol e paisagens bucólicas. Dessa maneira, o artista se mudou para o sul da França. Essa mudança fez com que Van Gogh trabalhasse, com entusiasmo, pintando flores e barcos com um estilo mais solto e expressivo.

Com um plano de criar uma 'colônia dos artistas', na cidade de Arles, Van Gogh

(a) (b)

Figura 2.2: (a) Vincent van Gogh - The Potato Eaters 1885 - (b) Vincent Van Gogh - Skull of a Skeleton with Burning Cigarette 1886 [1].

alugou quatro quartos. Um desses quartos, teve como inquilino o artista Paul Gauguin. Os dois trabalharam juntos e desenvolveram pinturas excepcionais, entretanto, tinham visões bastantes distintas sobre arte, analisando que, enquanto Paul Gauguin trabalhava com a imaginação, Van Gogh preferia pintar cenas reais. Tal dissemelhança entre as opiniões, acarretou diversas brigas e desentendimentos. Entre elas, Paul Gauguin ameaçou sair e Vincent chegou a ameaçá-lo com uma navalha. Nesta noite, o artista cortou a própria orelha e presenteou uma prostituta no distrito.

Após esse acontecimento, Van Gogh foi internado no Hospital de Arles. Segundo as cartas de seu irmão, ele estava mostrando sintomas daquela doença mais terrível, a loucura. Naquele tempo, ainda não existia toda uma teoria de psicanálise e muito menos diagnósticos de doenças como depressão, bipolaridade e outros distúrbios psicológicos, que provavelmente estiveram presentes durante toda a vida do pintor.

Nos meses seguintes, mesmo com a alta do hospital, Vincent Van Gogh continuava apresentando uma saúde mental perturbada e ainda assim, seguia adiante com suas atividades artísticas. Temendo um novo surto da doença, ele próprio internou-se no Hospital Psiquiátrico Saint Paulo de Mausole, em Saint-Rémy. Quando ainda estava internado, Van Gogh começou a trabalhar novamente, pintando paisagens internas e externas ao local. Em menos de um ano, o pintor holandês produziu cerca de 150 pinturas. Entretanto, sua saúde mental apresentava constante flutuação. Um exemplo deste estado de Van Gogh é o fato, relatado pelo hospital, de que ele havia comido tinta a óleo nesse período dando grandes indícios de 'loucura'.

Em julho de 1890, Vincent Van Gogh morreu com um tiro no peito, após receber uma breve alta do hospital psiquiátrico, e até hoje não se sabe ao certo se foi o próprio artista

que efetuou o disparo ou um terceiro. Dessa forma, o artista encerrou sua curta carreira como pintor resultando em apenas nove anos no total. Van Gogh não conseguiu vender um quadro quando vivo e teve sua saúde mental duramente prejudicada no final da vida, devido à falta de um diagnóstico. No entanto, ao longo do tempo foi reconhecido como um dos maiores pintores da cultura ocidental.

Figura 2.3: Self Portrait with Bandaged Ear, 1889 [1].

Capítulo 3

Fundamentação Teórica

Este capítulo introduz os conceitos computacionais utilizados para a realização da metodologia que será apresentada em seguida.

3.1 Processamento de Linguagem Natural

A comunicação natural entre o ser humano e a máquina está entre os paradigmas da computação desde seus primórdios [8]. Como no Teste de Turing, em 1950, que é baseado na comunicação em linguagem natural entre o ser humano e a máquina [9].

Com o avanço do poder de processamento e dos algoritmos de Inteligência Artificial, originou-se uma área denominada Processamento de Linguagem Natural que está presente em nosso cotidiano em diversas aplicações, como: assistentes de voz, tradutores de textos e robôs que realizam o atendimento de forma automatizada [10].

Podemos definir processamento de linguagem natural como a intersecção de três áreas: a Ciência da Computação, a Linguística e a Inteligência Artificial [13]. Seu principal objetivo é a compreensão e entendimento da linguagem para execução de tarefas por parte da máquina [8].

Do ponto de vista linguístico, o processamento de linguagem natural atua em dois níveis: o sintático e o semântico [12] . Na parte sintática, analisa-se a função e representação de cada palavra em um corpo de texto, e na parte semântica, considera-se a relação e interpretação entre as palavras e sentenças.

Com o advento da Inteligência rtificial e da computação foi possível realizar análises e automatizações, não somente para a execução de tarefas mas sim para análises mais profundas como análise de sentimentos e emoções.

Para usufruir de todas essas vantagens do processamento de linguagem natural, seguimos uma metodologia clássica na maioria dos trabalhos, desde da extração de informações até na aplicação mais robusta de algoritmos de aprendizado de máquina.

Portanto esse capítulo visa definir as etapas de: aquisição de dados, remoção de *stopwords*, modelo *bag-of-words*, stematização, lematização, identificação das classes gramaticais, a biblioteca NLTK e o modelo Word2Vec. Além de conceitos importantes para o maior entendimento do método proposto.

3.1.1 Linguagem Natural

A língua natural pode ser definida como uma série de caracteres, disponíveis em um conjunto bem definido de símbolos (alfabeto), que possuem como intuito a comunicação entre dois ou mais agentes. Os elementos básicos de qualquer língua são as palavras, caracterizadas como um agrupamento de caracteres que possuem um significado bem definido; denominamos um conjunto de palavras que detém algum sentido agregado, como uma sentença ou oração [10].

Toda língua formal possui uma gramática, no qual se estabelece um conjunto de regras para o uso da mesma. Sua estrutura é formada pela parte léxica, que especifica os critérios de formação das palavras, assim como suas classes morfológicas sendo: nomes, pronomes, adjetivos, proposições, conjunções e advérbios. Há também normas que definem funções e significados entre palavras e sentenças, sendo essas encontradas na área da análise semântica [10].

Outro fator importante a se destacar na linguagem natural é a ambiguidade. As variações da língua permitem que uma sentença possa ter mais de um significado sendo até mesmo no sentido oposto daquele explicitado, dependendo da sua pontuação ou, no caso da fala, da intonação aplicada a mesma. Essa dificuldade de interpretação do real significado das sentenças é o que torna o processamento da linguagem natural extremamente difícil e, da mesma maneira uma atraente área da computação e da comunidade científica.

3.1.2 Método de Aquisição de Dados da Web

Usualmente, utiliza-se para algoritmos de Aprendizado de Máquina [8] um conjunto de dados disponíveis em grandes bases públicas de forma bem estruturada, ou seja, com atributos bem definidos e muitas vezes rótulos já determinados. No caso de um domínio que ainda não foi explorado, faz necessária a coleta dos mesmos.

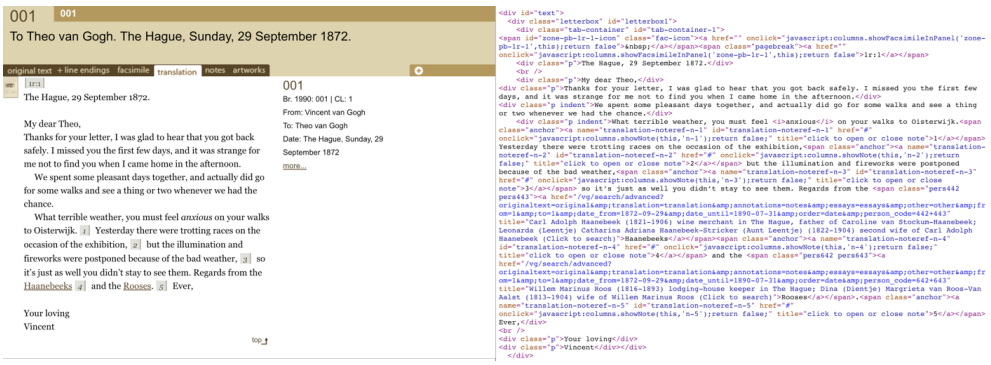
Atualmente a maior base de dados textuais está contida na Internet, porém, de forma totalmente desestruturada, desorganizada e contendo muitos ruídos [12]. Logo, é necessário aplicar métodos para a coleta desses dados de forma repetitiva e de maneira que se construa, ao final, um conjunto de dados robusto e de fácil manuseio.

Para obter os dados do presente trabalho, empregaram-se técnicas de webScraping, que consistem justamente em dado um conjunto de páginas HTML(sítios na internet) utilizar-se de suas estruturas de marcação para obter os dados de forma estruturada [11].

As etapas para execução desse processo são:

- *Download* dos sítios HTML
- Transformação e Normalização da informação
- Armazenamento dos dados
- Repetir o processo para sítios ou diretóeios similares

Com esse processo, é possível coletar um grande volume de dados de forma estruturada e organizada [11]. Na fase de transformação e normalização, faz-se necessário um processamento do texto adquirido, devido à existência de caracteres não desejados como: as tags HTML, números de marcação e espaços desnecessários muitas vezes contidos em sítios para melhora de estética e disposição no *browser*. Exemplificado, em linguagem Python, na Figura 3.1.



(a)

```
from bs4 import BeautifulSoup as BS
import requests

link = 'http://vangoghletters.org/vg/letters/let001/letter.html'
html = BS(requests.get(link).content, 'html.parser')
text = html.find_all(attrs={'id':'letterbox1'})[0]
```

```
text.getText()
```

```
'\n\n\xa01r:1\nThe Hague, 29 September 1872.\n\nMy dear Theo,\nThanks for your letter, I was glad to hear that you go
t back safely. I missed you the first few days, and it was strange for me not to find you when I came home in the aft
ernoon.\nWe spent some pleasant days together, and actually did go for some walks and see a thing or two whenever we
had the chance.\nWhat terrible weather, you must feel anxious on your walks to Oisterwijk.1 Yesterday there were trot
ting races on the occasion of the exhibition,2 but the illumination and fireworks were postponed because of the bad w
eather,3 so it's just as well you didn't stay to see them. Regards from the Haanebeeks4 and the Rooses.5 Ever,\n\nYou
r loving\nVincent\n'
```

(b)

Figura 3.1: (a) Exemplifica o sítio do Museu Van Gogh, a direita a disposição do browser e a esquerda o código HTML (b) Código de extração do texto do sítio presente em (a)

Tabela 3.1: Frequência do trecho no formato original.

Palavras	Frequência	Palavras após Limpeza	Frequência após Limpeza
vincent	5	vincent	5
and	4	gogh	4
the	4	van	4
van	4	dream	2
gogh	4	new	2

3.1.3 Limpeza de Dados: Remoção de Stopwords

Um documento é formado por um número de palavras únicas, sendo essas definidas como seu vocabulário. Os vocábulos mais frequentes, são muitas das vezes conjunções, proposições, pronomes e artigos que estão presentes para manter a estrutura semântica e de coerência do texto, porém, não possuem nenhum significado embutido. Essas classes de palavras são denominadas *stopwords* [12, 13].

A remoção dessas palavras faz-se necessária uma vez que atrapalham o desempenho em análises futuras, já que não possuem nenhum siginificado semanticamente. Para o melhor entendimento, temos um trecho retirado do sítio do Museu Van Gogh sobre sua passagem pela cidade de Paris e sua ida ao sul da França:

> *"In 1888 Vincent van Gogh leaves Paris behind and travels to Arles. The big city took a heavy toll on Vincent's health. In the sunny south of France, Vincent hopes to find peace and quiet. Blisfully happy with his new surroundings, Van Gogh works tirelessly towards realising his dream: creating a new kind of art full of light and colour, together with other artisits. However, things turn out differently and Vincent's dream falls apart. Van Gogh Dreams tells the story of this turbulent period in Van Gogh's life across five different areas. It is based on the many letters that Vincent wrote to his brother Theo."*[1]

O trecho possui inúmeras palavras que não agregam sentindo ao texto (*stopwords*), logo é realizado a sua remoção para uma comparação. A Tabela 3.1 mostra as cinco palavras únicas que mais aparecem no trecho citado. Após a remoção, obtemos a Tabela 3.1, na qual as palavras únicas mais frequentes desse novo trecho contém apenas vocábulos com significados expressivos para uma análise sintática ou semântica futura.

> *"1888 vincent van gogh leaves paris travels arles big city took heavy toll vincent health sunny south france vincent hopes peace quiet blisfully happy new surroundings van gogh works tirelessly realising dream creating new kind art light colour artisits things turn differently vincent dream falls apart van gogh dreams tells story turbulent period van gogh life different areas based letters vincent wrote brother theo"*

[1] `http://www.vangoghmuseum.nl/en/whats-on/exhibitions/van-gogh-dreams`

3.1.4 Modelo Bag of Words

O modelo *bag of words* é o método de representação textual baseado na frequência da unidade básica do texto [10], no caso cada palavra. Sendo sua entrada o corpo do texto, D, esse formado por um conjunto de palavras, x_i, podendo ser representado por:

$$D = x_1, x_2, x_3, x_4, x_5...x_n \tag{3.1}$$

Com essa entrada, estabelecemos que cada palavra deve ser representada por apenas um elemento x_n, sem repetições em nossa representação e não considerando sua posição de ocorrência. Definimos esse processo como tokenização do texto [10].

Com o vetor de entrada definido, realiza-se a contagem da ocorrência de cada palavra no corpo do texto, resultando em um vetor de saída que determina a frequência de cada palavra determinado pelo índice, que pode ser definido por:

$$F = f_1, f_2, f_3, f_4, f_5...f_n \tag{3.2}$$

O método descrito acima gera uma representação de forma geral para um determinado texto de entrada, mostrando a importância de cada palavra no corpo do texto [14, 10]. Para representação visual desse modelo, usa-se a nuvem de palavras, no qual se determina um gráfico de palavras em que quanto maior a frequência maior será o tamanho da sua fonte e vice, e versa, sendo possível visualizar as palavras-chaves de uma entrada de maneira fácil e intuitiva.

Tendo como exemplo o trecho retirado do sítio do Museu Van Gogh abaixo, podemos visualizar frequência pela nuvem de palavras em Figura 3.2.

> *"Extensive research has been conducted on Sunflowers in recent years. We now know more about the fascinating genesis and condition of the work. The exhibition Van Gogh and the Sunflowers allows visitors to discover how important the sunflower was to Van Gogh and learn the answers to pressing questions such as how we can best preserve this painting so that it can be enjoyed by generations to come."*[2]

Uma extensão do modelo *bag of words*, é a frequências dos n-grams que são sequências de palavras que aparecem muitas vezes juntas ao longo do texto. O modelo descrito acima, por exemplo, exibe o n-grama com $n = 1$, ou seja, apenas verifica a ocorrência da palavra. Já modelos com $n > 1$, temos uma probabilidade de ocorrência de nomes compostos, sentenças e orações ao longo do corpo textual.

Para o texto utilizado no primeiro exemplo dessa seção, podemos ver na Tabela 3.2 os n-gramas formados com $n = 2$, denominados bigramas.

[2] `http://vangoghmuseum.nl/en/whats-on/exhibitions/exhibition-van-gogh-and-the-sunflowers`

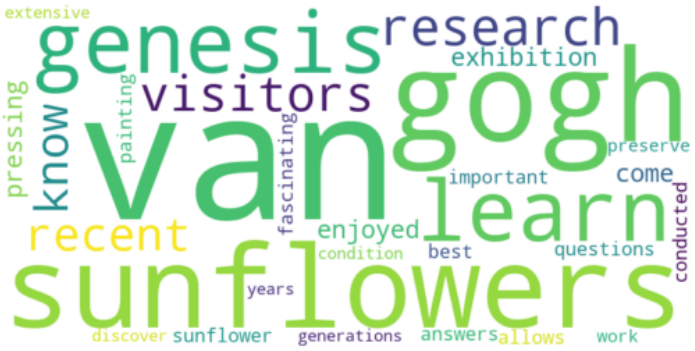


Figura 3.2: Exemplo da nuvem de palavras do trecho retirado do Museu Van Gogh

Tabela 3.2: Exemplo de bigramas do fragmento de texto retirado do Museu Van Gogh

N-Grama	**Frequência**
extensive research	1
research has	1
has been	1
been conducted	1
conducted on	1
on Sunflowers	1
Sunflowers in	1
in recent	1
recent years.	1

3.1.5 Stemização

Em um conjunto de documentos, as palavras sofrem flexões e variações de acordo com sua classe gramatical. Desse modo, ocorrem diversas mudanças em sua grafia com a adição de sufixos e prefixos ou até mesmo mudanças mais abruptas para fornecer um determinado grau, ou gênero a um vocábulo.

O processo de stematização é aquele no qual o radical das palavras, seu sentido básico e indivisível, é identificado [15]. Para realizar esse procedimento, é utilizam-se de expressões regulares em conjunto com operações de divisão de *strings* para a extração apenas do radical, removendo das palavras os sufixos ou prefixos. Na stematização, o radical resultante não precisa ter um significado, ou seja, essa técnica apenas permite identificar, nas mais diversas formas linguísticas apresentadas, o morfema básico daquelas palavras [12].

Esse método é utilizado para melhorar o desempenho dos modelos de aquisição de características *bag-of-words* com o intuito de identificar a unicidade das palavras em um vocabulário, na análise de sentimentos e objetividade entre outras operações que envolvem a contagem ou amostragem de palavras.

3.1.6 Lematização

Lematização é o processo que mapeia a forma na qual uma palavra aparece, em uma unidade básica com significado, ou seja, é feito um processo parecido com a stematização, porém, mapeando sempre para um radical que constitui uma palavra ou morfema com um significado semântico [14].

Esse radical com significado é conhecido como lema, que é a forma gramatical no qual se representa um morfema ou palavras em suas mais variadas formas. Como exemplo, tendo o verbo **ser** e suas conjugações em diversos tempos verbais: presente, pretérito perfeito e futuro do presente no indicativo (descrito na Tabela 3.3). Essas diversas formas representam um mesmo lema, que seria a palavra **ser**.

Assim como na Língua Portuguêsa, podemos aplicar o mesmo processo com variações linguísticas de outras línguas como na Língua Inglesa. Considerando a palavra **sing**, temos suas variações como *sang* e *sung*, porém, seu lema é centrada na palavra **sing**, que traduzindo livremente para o português seria o verbo cantar [13].

3.1.7 Identificação de Classe Gramatical

A ambiguidade das palavras em seus diversos contextos são considerados uma grande dificuldade no processamento de linguagem natural. Visto que a língua natural possui

Tabela 3.3: Flexibilização do verbo ser, exemplificado pela sua conjugação

Pessoa	Presente	Pretérito Perfeito	Futuro do Presente
eu	sou	fui	serei
tu	és	foste	serás
ele	é	foi	será
nós	somos	fomos	seremos
vós	sois	fostes	sereis
eles	são	foram	serão

diversas formas de expressão para um determinado assunto, utilizam-se técnicas para diminuir ao máximo a ambiguidade e o entendimento errôneo por parte da máquina.

Portanto, aplica-se a identificação de classes gramaticas ou *part-of-speech*, para determinar a classe sintática, na qual cada palavra em nosso corpo de textos pertence. Para que esse processo ocorra, deve-se remover a pontuação do conteúdo a ser analisado e, além disso, transformá-lo em uma lista de palavras para conseguirmos associar sua classe gramatical [15].

Um exemplo das classes sintáticas são: nomes, pronomes, verbos, conjugação, advérbio, adjetivo, entre outros. Esse mapeamento é feito utilizando regras pré-definidas a partir de um conjunto de dados contendo uma série de variações da linguagem e regras gramaticais com o intuito de definir quando uma palavra se comporta como um adjetivo ou um nome, por exemplo; ou utilizando *part-of-speech* estocástico no qual é feito um treinamento prévio com os dados textuais para calcular a probabilidade de um determinada palavra ser de uma classe dado aquele contexto. Esse método utiliza algum modelo pré-definido para sua execução, como, por exemplo, Modelo Oculto de Markov ou *part-of-speech* Brill [14].

Ao final desse processo temos dado uma sentença de entrada qual função sintática cada palavra está exercendo, o que facilita em análises futuras por retirar em parte a ambiguidade e atribuir uma função para cada elemento contido naquele escopo.

Figura 3.3: Exemplo de *part-of-speech* da frase: "Van Gogh loves to paint at afternoon.". Em vermelho os verbos, em cinza os substantivos, em amarelo as preposições e preto estruturas auxiliares.

3.1.8 Natural Language Toolkit

Natural Language Toolkit, NLTK, consiste em uma biblioteca, escrita em linguagem de programação Python, que possui diversas interfaces utilizadas para processamento de

linguagem natural. Desenvolvido no Departamento de Computação da Universidade da Pennsylvania em 2001, passou a ser o principal framework com implementações robustas das principais ferramentas de área [16].

Seus principais módulos incluem: banco de dados com diversos textos para testes, funções para processamento de *string* (conjunto de caracteres), funções para coleta de material descritivo de um dado texto, identificação de classe gramatical, expressões regulares, *parser*, métricas de avaliação, dicionários e aplicações para desenvolvimento de chats [16].

Dessa forma, empregam-se suas funções para um rápido desenvolvimento e teste do *pipeline* de processamento de linguagem natural escolhido determinado domínio problema.

3.1.9 Análise de Sentimentos e Extração de Emoções

Com o avanço da área de processamento de linguagem natural, surgiu um campo que tem como objetivo identificar as emoções presentes e a polaridade de textos. Essas análises também pertencem ao paradigma da interação entre máquina e o ser humano na execução de tarefas.

O conceito de emoção é bastante complexo no ramo da psicologia, porém, sempre está fortemente relacionado com sentimentos, comportamentos, mudanças fisiológicas e cognitivas, além de sempre ocorrer em um contexto particular influenciando todos os outros conceitos. Ou seja, sua função principal é a relação entre o indivíduo e sua interação com o mundo [2].

O psicólogo Plutchik propôs em sua teoria uma abordagem de classificação para determinadas emoções fundamentais: raiva, medo, tristeza, nojo, surpresa, antecipação, confiança, e felicidade. Ele considerava que cada uma dessas emoções primárias tinha influências no comportamento de sobrevivência do ser humano.

Além disso, Plutchik [2] elaborou o círculo das emoções que define as emoções secundárias e terciarias além da bipolaridade entre elas, podemos visualizar na Figura 3.4 de forma gráfica esta representação. Uma expansão dessa teoria relaciona as emoções fundamentais entre si para definir um sentimento, essas combinações são denominadas dyads. Na Tabela 3.4 exemplifica-se os *dyads*.

A partir da teoria do campo da psicologia, podemos dividir os métodos para a extração e identificação de emoções nos textos em 3 abordagens: palavras-chave, Aprendizado de Máquina e híbrido.

Os métodos baseados em palavras chaves utilizam as mesmas para detectar as emoções nos textos. Utiliza-se muitas vezes um conjunto de sinônimos para facilitar de identificação das mesmas. Geralmente usa-se o auxílio de um dicionário léxico como a WordNet para diminuir o risco de ambiguidade e orientação linguística.

Tabela 3.4: Dyads - Combinações das Emoções. [2]

Sentimentos	**Emoções**	**Sentimentos Opostos**	**Emoções**
Optimism	Anticipation + Joy	Disapproval	Surprise + Sadness
Hope	Anticipation + Trust	Unbelief	Surprise + Disgust
Anxiety	Anticipation + Fear	Outrage	Surprise + Anger
Love	Joy + Trust	Remorse	Sadness + Disgust
Guilt	Joy + Fear	Envy	Sadness + Anger
Delight	Joy + Surprise	Pessimism	Sadness + Anticipation
Submission	Trust + Fear	Contempt	Disgust + Anger
Curiosity	Trust + Surprise	Cynicism	Disgust + Anticipation
Sentimentality	Trust + Sadness	Morbidness	Disgust + Joy
Awe	Fear + Surprise	Aggressiveness	Anger + Anticipation
Despair	Fear + Sadness	Pride	Anger + Joy
Shame	Fear + Disgust	Dominance	Anger + Trust

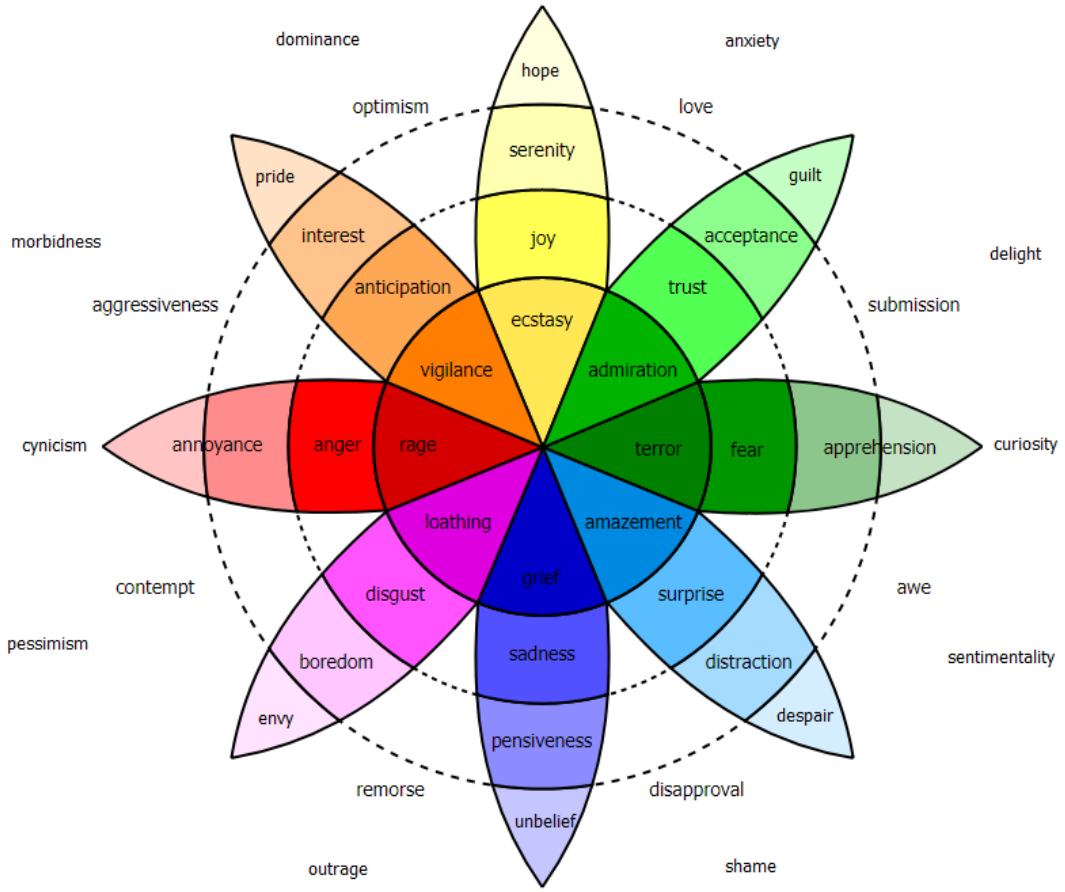


Figura 3.4: Círculo das Emoções definidos por Plutchik [2]

Já na abordagem que utiliza Aprendizado de Máquina, temos que a partir de um conjunto de dados textuais, já rotulados com cada emoção aplicar um modelo de classificação de múltiplas classes com o objetivo de melhorar o desempenho a cada iteração. Esse método geralmente utiliza classificadores como SVM e Naive Bayes, que lidam bem com representações textuais.

Como o uso desses dois últimos métodos apresentados são restritos por tamanho de texto, ambiguidade, dificuldade em uma pré-classificação e desempenho não satisfatório, surgiu a metodologia híbrida, a qual utiliza características das palavras chaves para alimentar o modelo de classificação com o intuito de melhorar sua precisão [17].

A extração de emoções e sentimentos de textos ainda é uma tarefa complexa e possui uma parte exploratória em cada domínio problema a ser analisado. Os métodos existentes indicam a existência e ocorrência daquela emoção ou sentimento em um determinado texto, por muitas vezes esbarrarem no grande problema da área de processamento de linguagem natural, a ambiguidade.

3.1.10 Modelo Word2Vec

Quando escolhemos analisar um texto somente utilizando uma abordagem léxica esbarramos em alguns problemas como: atualização constante do dicionário léxico visto que a linguagem natural é algo evolutivo, o contexto de certas palavras aplicadas em determinadas áreas, a subjetividade além de requerer esforço humano para criar e adaptar tais dicionários [18].

Dessa forma, buscou-se ultimamente a construção de um segundo tipo de abordagem que representa as palavras na forma de um vetor, no qual a ideia principal é que o significado da palavra é dado pelas palavras que aparecem frequentemente em sua vizinhança, essa correlação é denominado como contexto. Portanto, os vetores que refletem cada palavra devem ser próximos de palavras contidas em seus contextos [18].

O modelo *word2vec* foi criado baseado na abordagem citada anteriormente e tem como objetivo prever o contexto de uma uma palavra x_n [19].

Existem dois tipos de arquitetura: *skip-grams* e *bag of words* contínuo. O modelo skip-gram tem estrutura varrer todo o corpo do texto buscando por janelas de palavras próximas com o tamanho n e contendo uma palavra central. Exemplificando, supondo que temos a frase, após a remoção de stopwords:

"Vincent figuras famosas influentes história arte ocidental"

Logo, sendo a janela de contexto escolhida 1, teríamos os seguintes vetores:

$$C = [([Vincent, famosas], figuras), ([figuras, influentes], famosas),$$
$$([famosas, história], influentes), ([influentes, arte], história), ([história, ocidental], arte)], \tag{3.3}$$

Contendo sempre a estrutura de $([contexto],'palavracentral'))$, assim a tarefa do modelo é prever cada contexto dado como entrada uma palavra central, ou seja:

$$S_k(figuras) = [Vincent, famosas]. \tag{3.4}$$

Esse treinamento é feito com o auxílio de uma rede neural utilizando um classificador *softmax*, em que a partir de uma normalização do vetor de entrada, mapear sua saída em uma distribuição probabilística das classes de saída (no caso cada contexto):

$$p(w_o|x_I) = \frac{exp({v_{w_o}}'^{\top} v_{w_I})}{\sum_{w=1}^{W} exp({v_w}'^{\top} v_{w_I})}, \tag{3.5}$$

em que v_w e ${v_w}'$ são os vetores de entrada e saída, respectivamente, que representam cada palavra w, W o número de palavras em nosso vocabulário.

A outra abordagem, denominada bag of words contínuo realiza a operação inversa, ou seja, dado o contexto como entrada conseguirmos prever a palavra central, esse utiliza os mesmos métodos de treinamento e classificação previamente descritos.

Para exemplificar de maneira visual o modelo vetorial e as correlações de contexto de determinadas palavras temos a Figura 3.5

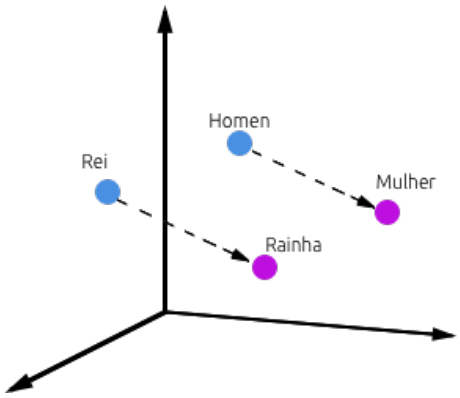


Figura 3.5: Exemplo Gráfico da Representação Vetorial de Palavras de um Texto

Como temos vetores simbolizando os vocábulos, conseguimos realizar operações aritméticas com cada palavra, encontrar similaridades em seus contextos e proximidades interessante, como exemplificado na

$$vec(\text{"}rei\text{"}) = vec(\text{"}rainha\text{"}) - vec(\text{"}homem\text{"}) \tag{3.6}$$

3.2 Imagens Digitais

Atualmente temos mais contato com câmeras em relação às décadas passadas, logo as imagens digitais se tornaram muito comuns em diversas aplicações como, por exemplo, em redes sociais e na digitalização de arte. Como a aquisição das imagens se tornou um procedimento mais fácil e ágil, houve um avanço na área de processamento de imagens.

Na fase de aquisição, as imagens digitais são obtidas através da captura da luz refletida pelos objetos que por meio de transformações não lineares produzem uma matriz constituída por uma série de elementos únicos denominados pixels, que representam a intensidade de cor e luz em uma cena, como definido abaixo [3]:

$$I = f(x, y), \tag{3.7}$$

O processo de aquisição pode ser exemplificado pela Figura 3.6.

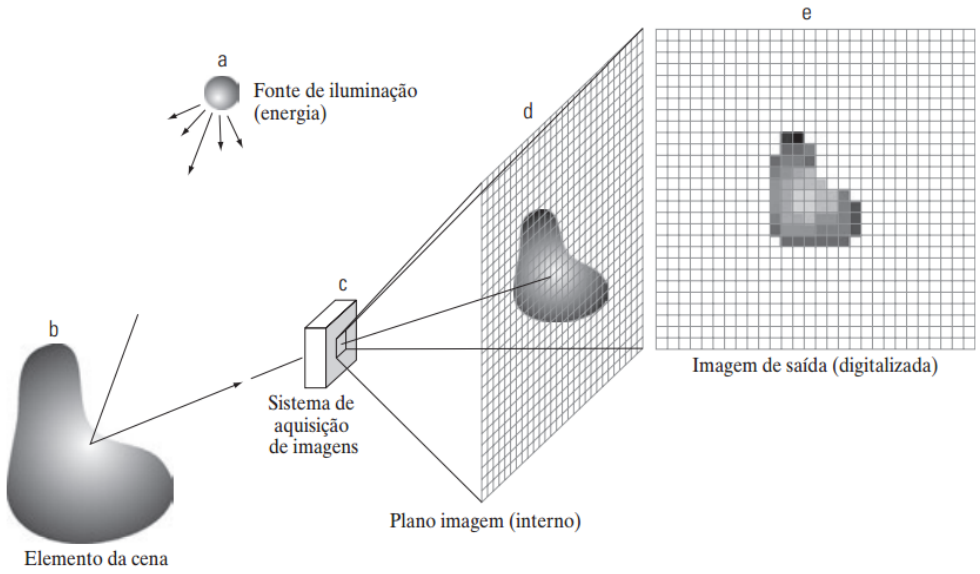


Figura 3.6: Processo de Aquisição de Imagens [3]

3.2.1 Nível de cinza e Imagens Coloridas

Existem duas categorias de imagens: em escala de cinza e coloridas. Imagens em escala de cinza são imagens representadas por uma matriz de uma dimensão onde cada pixel

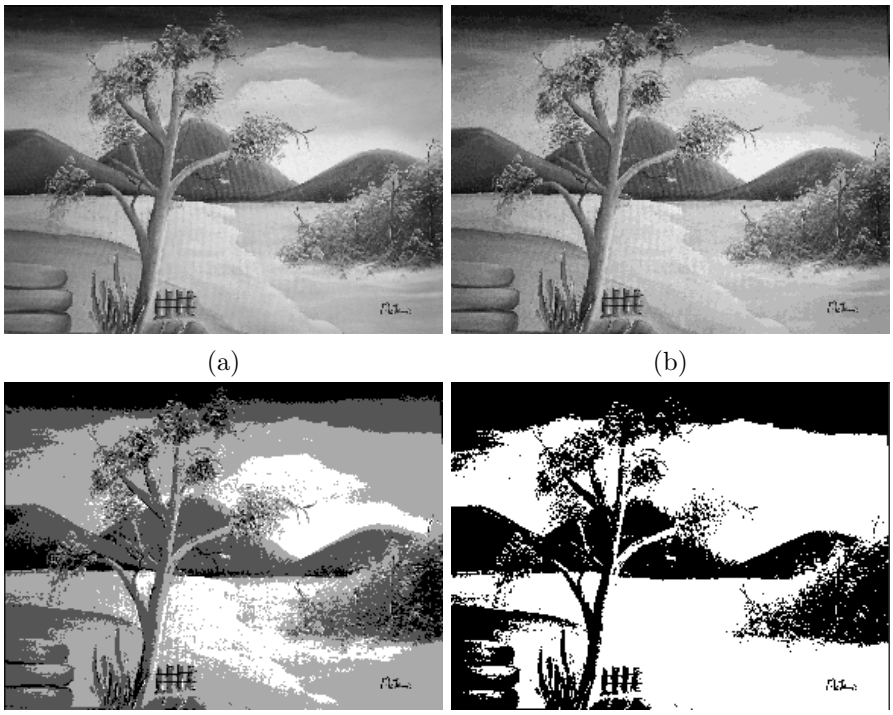
contém a intensidade de cinza da cena, a quantidade de níveis existentes varia de acordo com a quantidade disponível de bits para sua representação [3].

Quanto maior for a quantidade de bits teremos maiores detalhes em nossas imagens, pois, teremos uma cena melhor representada. Por exemplo, se temos 8 bits então dispomos de 2^8 níveis de cinza, ou seja, 256 tons (entre 0 e 255) na escala de cinza. Essa relação do número de níveis L é definida como:

$$L = 2^n, \tag{3.8}$$

sendo n, o número de bits disponíveis para representação do canal e L o número de níveis possíveis para cada pixel da imagem.

Quando definimos a quantidade dos nossos níveis de cinza estamos realizando a quantização da nossa imagem. Exemplificamos a variação dos níveis de cinza em Figura 3.7.



(a) (b)

(c) (d)

Figura 3.7: Quantização - (a) pixels de 8 bits e 256 níveis de cinza; (b) pixels de 4 bits e 16 níveis de cinza; (c) pixels de 2 bits e 4 níveis de cinza; e (d) pixels de 1 bit e 2 níveis de cinza.

Uma imagem colorida no espaço de cores RGB é formada por três componentes espectrais primários de vermelho, verde e azul. Cada componente contribui para formação da cor visualizada na imagem final, assim cada pixel em sua dimensão possui sua intensidade

de cor. Outros exemplos de espaços de representação de cores são o YCbCr e o HSL, que apresentam componentes de luminância e saturação.

(a) (b)

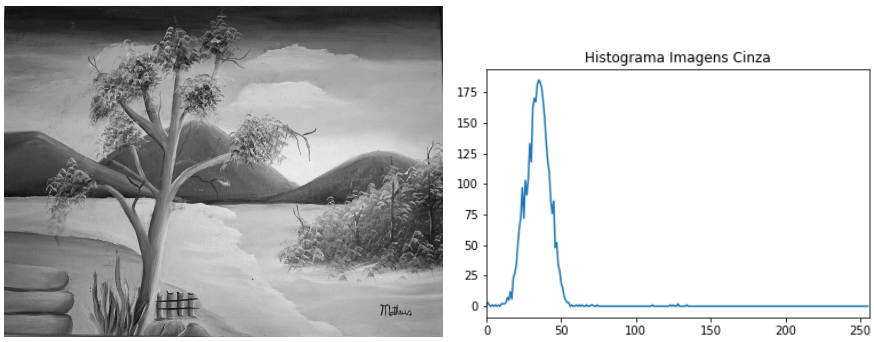
Figura 3.8: (a) Imagem Colorida (b) Imagem Escala de Cinza

3.2.2 Histograma

O histograma de uma imagem digital com níveis de intensidade no intervalo [0, L – 1] é uma função definida por:

$$h(r_k) = n_k, \tag{3.9}$$

sendo r_k é o k-ésimo valor de intensidade e n_k é o número de pixels da imagem com intensidade r_k, ou seja, sua frequência [3].

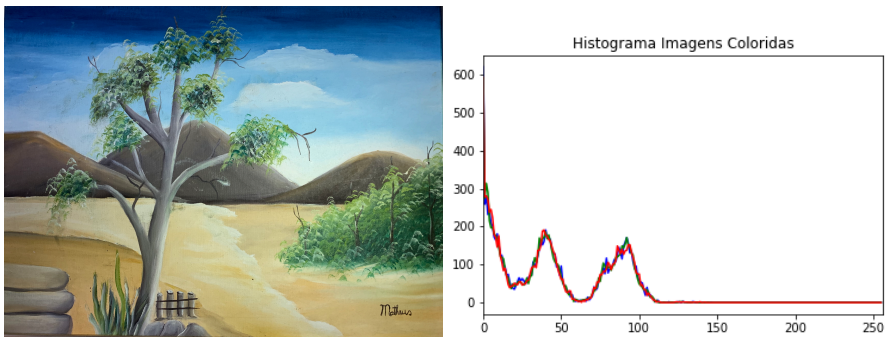


(a) (b)

Figura 3.9: Exemplo de histograma de uma imagem em nível de cinza

O histograma de uma imagem colorida no sistema de cores RGB é composto pela frequência dos pixels em cada componente de cor, resultando assim em 3 histogramas diferentes, sendo sua análise é feita de forma conjunta.

Histogramas são a base para várias técnicas de processamento no domínio espacial. Sua manipulação pode utilizada para realce e correção de imagem além de fornecer estatísticas úteis sobre a mesma.



(a) (b)

Figura 3.10: Exemplo de histograma de uma imagem colorida

3.2.3 Sistemas de Cores

O ser humano consegue visualizar e discernir milhares de tons e cores. Apesar de ser um processo do cérebro humano ainda não completamente definido, o processo fisiopsicológico, as cores podem ser expressas a partir de experimentos físicos com a luz [3].

Os modelos de cores são responsáveis por unificar a representação das cores de forma que possamos representar em um espaço de coordenadas cores dos mais diversos tons e intensidades.

Modelo de Cor RGB

O modelo de cor RGB é um modelo orientado ao hardware, ou seja, sua implementação busca uma maior proximidade com os monitores coloridos disponíveis no nosso dia a dia.

Nesse modelo, cada cor é formada por três componentes espectrais primários: vermelho, verde e azul [3]. Esse baseia-se em modelo de coordenadas cartesianas e o subespaço de interesse consiste em cubo como representado em Figura 3.11.

Cada vértice primário x, y e z representa uma cor primária: vermelho, verde e azul. As cores secundárias são representadas pelas extremidades do cubo: magenta, ciano e amarelo. Além de tanto a origem como o ponto mais distante do cubo simbolizam o preto e o branco respectivamente.

As outras cores estão espalhadas por pontos dentro desse subespaço, onde os valores são normalizados de 0 até 1 de forma que a representação de bits escolhidas apenas seja uma fator de profundidade, alterando assim o número de cores disponíveis em nosso

modelo. Como o modelo possui três componentes espaciais se temos 8 bits para cada canal, logicamente possuímos o total de $(2^8)^3 = 16.177.216$ cores.

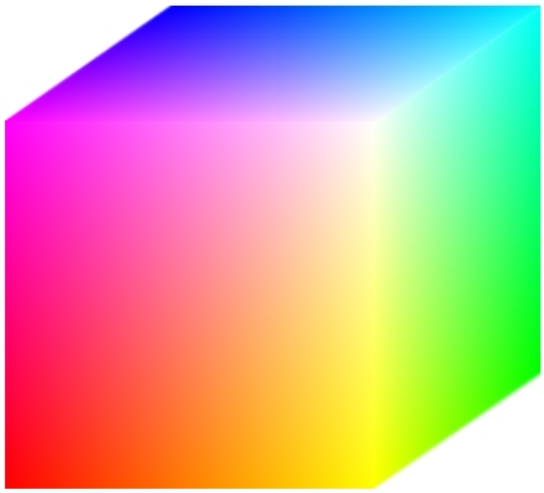


Figura 3.11: Cubo de cores do modelo RGB. Retirado de [3]

Modelo de Cor CIEL*a*b

O modelo de cores LAB ou CIEL*a*b foi proposto pela Comissão Internacional de Iluminação em 1976 e explicita suas cores em três componentes, sendo elas: o L que determina a intensidade da luminosidade, ou seja, 0 sendo preto e 100 sendo branco absoluto; o *a que define a intensidade entre os limiares verde e vermelho; e *b que representa o intervalo de intensidades entre azul e o amarelo[3].

Por cada componente ser representado por números reais, o número de cores que conseguimos simbolizar com esse modelo é praticamente infinito. Outra questão abordada é a relação não linear das componentes o que permite calcular a distância entre cores de forma mais robusta.

No modelo CIEL*a*b a distância entre dois pontos no espaço determina a diferença entre duas cores nos critérios de luminância, croma e matiz, portanto, é geralmente utilizado para calcular a similaridade entre duas cores pelo padrão determinado pelo CIEL*a*b:

$$\Delta E = \sqrt{(L_2 - L_1)^2 + (*a_2 - *a_1)^2 + (*b_2 - *b_1)^2}. \tag{3.10}$$

3.2.4 Matriz de Co-ocorrência de Níveis de Cinza

Matriz de Co-ocorrência de Níveis de Cinza, (GLCM, Gray-Level Co-occurrence Matrix), é o método no qual são extraídas estatísticas da imagem pela caracterização de sua textura

através da criação de uma matriz que indica quantas vezes determinado par de pixel em uma certa direção [4].

O algoritmo funciona varrendo a nossa imagem escolhendo um determinado par de pixels, I(1,1) e I(1,2), contando quantas vezes o segundo pixel ocorre em seguida do primeiro pixel, como demonstrado na Figura 3.12. Em uma próxima iteração, o algoritmo varia a direção da ocorrência entre 45 e 90 graus, essa contagem é armazenada em uma nova matriz no qual são retiradas as características de: energia, correlação, contraste e homogeneidade.

A *Energia* representa a uniformidade da textura e é calculada por:

$$Energia = \sum_{i,j} GLCM(i,j)^2. \tag{3.11}$$

O *Contraste* refere-se a diferença entre o pixel e todos os outros da imagem e é obtido a partir de:

$$Contraste = \sum_{i,j} |i-j|^2 GLCM(i,j). \tag{3.12}$$

Já a *Correlação* define o quão correlacionado o pixel está com sua vizinhaça, sendo essa:

$$Correlação = \sum_{i,j} \frac{(i-ui)(j-uj)GLCM(i,j)}{\sigma_i \sigma_j}. \tag{3.13}$$

E por fim, temos a *Homogeniedade* que mede a proximidade das distribuições dos elementos em relação a diagonal da matriz $GLCM$, calculado por:

$$Homogeniedade = \sum_{i,j} \frac{GLCM(i,j)}{1+|i-j|} \tag{3.14}$$

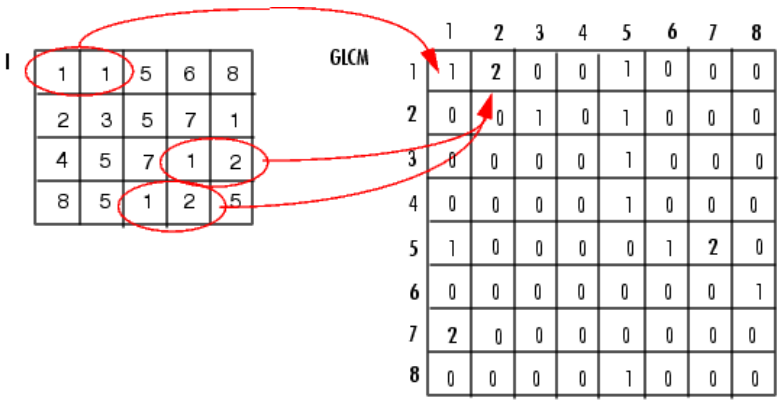


Figura 3.12: Exemplo da construção da matriz de co-ocorrência de níveis de cinza [4]

3.3 Aprendizado de Máquina

Aprendizado de Máquina, de modo simples, é como a partir de um domínio problema chegar ao domínio solução, através de um treinamento prévio com exemplos variáveis de um conjunto desses dois domínios. Os algoritmos aplicados são capazes de realizar basicamente dois tipos de tarefas, a classificação de padrões e a previsão de valores.

A classificação de padrões é definida como tendo um conjunto de dados com n padrões estabelecidos. Ppodemos a partir de uma nova entrada enquadrá-la em um dos padrões com a maior probabilidade de semelhança [8, 20]. Os padrões são definidos a partir de um conjugado de características relevantes que os delimitam.

Já previsão de valores refere-se a geração de novas informações baseando-se em um conjunto de exemplos prévios. Seus algoritmos e métodos têm como foco a entrega de um resultado numérico e único e não de classes como na classificação.

Todo modelo desenvolvido em aprendizado de máquina segue um conjunto de tarefas que são categorizadas em 5 etapas principais (Figura 3.13): coleta da base de dados, entendimento do problema, limpeza e normalização da nossa base, treinamento do modelo e validação do mesmo [5].

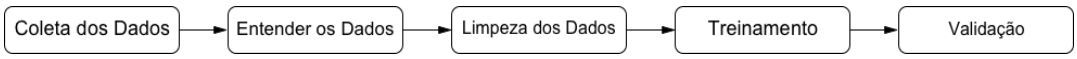


Figura 3.13: Pipeline Aprendizado de Máquina

A etapa de coleta de dados refere-se a aquisição dos exemplos do domínio problema uma vez que os algoritmos conhecidos exigem um grande volume dos mesmos em nossa base. O entendimento dos mesmos passa pelo conhecimento do problema a ser resolvido e de que forma pode ser solucionado pelo conjunto de características disponíveis.

A limpeza da nossa base é um processo de extrema importância para os algoritmos de aprendizagem de máquina uma vez que na coleta dos dados pode ter ocorrido algum erro de medição ou humano, a falta de determinada característica e até mesmo a presença de exemplos duplicados não resultando em um aprendizado eficaz pelo algoritmo. De modo que a limpeza serve para evitar duas fraquezas dos algoritmos de aprendizado de máquina: o bias e a variância, que são a tendência do modelo de aprender errado ou de forma aleatória, respectivamente. A limpeza pode influenciar também no *overfitting* que é o vício no treinamento do nosso modelo, que muitas vezes ocorre devido a não separação do conjunto de dados em treinamento, validação e teste. Ou seja, viciamos o nosso modelo a acertar as previsões de todos os nossos exemplos disponíveis, mesmo sabendo da diversidade do domínio solução, causando assim resultados com baixo desempenho na resolução da tarefa real.

Após a limpeza, devemos escolher qual algoritmo se encaixa melhor em um determinado problema. Os algoritmos disponíveis como redes neurais, máquinas vetoriais e árvores de decisão possuem suas vantagens e desvantagens para cada tipo de dado. Com esse cenário, a validação e teste se tornam uma importante fase para o *pipeline*, visto que somente nesta fase será avaliada de forma numérica qual modelo ofereceu o melhor desempenho ou até mesmo retornar ás fases anteriores para optimizar o mesmo e melhorar sua taxa de acerto ou previsão.

3.3.1 Classificação de Padrões Não Supervisionada

Para classificação de padrões de maneira não supervisionadas utiliza-se principalmente algoritmos de agrupamento. Esses tem como principal objetivo a partição de um conjunto de dados em subconjuntos, denominados *clusters*. Cada *cluster* possui objetos similares entre si, sendo que os menos semelhantes pertencem a outros *clusters* e assim por diante.

Os resultados do processo de agrupamento variam dado o algoritmo utilizado e até mesmo em suas iterações, podendo obter, ao final, resultados diferentes para um mesmo conjunto de dados [5].

Uma das principais vantagens desse tipo de abordagem é a descoberta de grupos no qual uma análise humana desconhece dado um conjunto de dados. Logo, tendo dados não rotulados pode-se definir padrões sem uma classificação prévia(treinamento).

Para o escopo do nosso trabalho, utilizou-se o algoritmo *K-means*.

O *K-means* é o algoritmo de agrupamento baseado na distância entre os exemplos. Supondo que nossos dados, D, estão contidos em n objetos no espaço Euclidiano, devemos dividi-los em um número de *clusters* desejáveis k, portanto se aplica uma função de custo com o objetivo de melhorar, a cada iteração, a similaridade entre objetos em um mesmo *cluster* e, ao mesmo tempo, a diferenciação entre objetos em *clusters* diferentes [5].

Como o método é baseado no ponto central de cada *cluster*, inicialmente é feita a distribuição dos centroides em que a cada iteração têm-se suas posições ajustadas de maneira que a diferença entre objetos e o centroide sejam determinadas pela distância euclidiana entre eles.

Ao final das iterações formam-se grupos baseados em cada centroide que possuem uma certa qualidade, ou seja, a função de custo chegou ao seu mínimo. Esse processo pode ser exemplificado na Figura 3.14.

Tendo em vista que a escolha do parâmetro k é o que define a quantidade de *clusters* pretendidos, utiliza-se o método de *elbow*. Esse é baseado na observação do conjunto de dados e sua variação dado diversos valores de k. A utilização do método previne a execução de k muito altos nos quais dois ou mais grupos não agregam ao ficarem separados

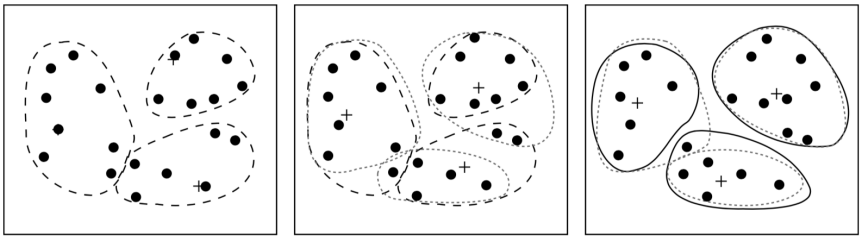


Figura 3.14: Processo de agrupamento da primeira a última iteração[5]

e assim por diante, da mesma forma com um k muito pequeno no qual não se consegue distinguir um grupo do outro.

Capítulo 4

Trabalhos Correlatos

Nesse capítulo iremos apresentar alguns trabalhos correlatos utilizados na elaboração da solução proposta como base teórica e experimental. Foram levadas em consideração as duas áreas pertinentes para a execução do presente trabalho, Processamento de Imagens com arte e Processamento de Linguagem Natural.

Em [21] é em agrupar *tweets* (mensagens curtas até 240 caracteres) de acordo com emoção representada pelo mesmo. Primeiramente, definem-se três categorias principais de sentimentos: positivos, neutros e negativos. O grupo positivo é caracterizado pelo uso de palavras que contêm sentido positivo para um determinado fato, da mesma forma o grupo negativo é aquele que possui palavras de cunho negativo geralmente envolvido em uma crítica a um acontecimento e no caso neutro seria o meio-termo de nenhum dos dois últimos grupos citados. Utilizam-se nesse trabalho algoritmos de aprendizado de máquina não supervisionados, como procedimentos de clusterização que têm como principal tarefa a definição de grupos com certas similaridades de maneira não supervisionada dado um conjunto de dados de entrada.

O método apresentado é dividido em (I) extração de dados, (II) limpeza dos dados, (III) tokenização, (IV) análise sentimental via *scores* e (V) aplicação dos algoritmos de agrupamento. Na seção de extração de dados foi utilizado a API disponibilizada pelo próprio Twitter para obtenção de 150 mil *tweets*. Com os dados coletados, foram realizada a limpeza de dados (II) para a exclusão de itens que não possuíam textos, irregularidades e inconsistências. A tokenização é o processo na qual se separa o corpo do texto em seu menor nível de informação, a palavra (token), esse método também é conhecido como analise léxica. Esse permiti aos algoritmos um entendimento melhor da linguagem. Após esse procedimento realizou-se a obtenção dos *scores* sentimentais de cada palavra utilizando duas bases de dados já treinadas sendo essas a Afinn e o TextBlob, o autor não deixa de forma clara o cálculo do *score* total de um *tweet* apenas de cada palavras.

Com as polaridades obtidas de cada *tweet*, foi aplicado o algoritmo *K-means*, no qual

a partir de inúmeras iterações consegue separar um conjunto de dados em k grupos, sendo k determinado previamente, com centroides iniciais que são recalculados a cada iteração para se adequar ao conjunto de dados, o k escolhido foi determinado pela curva de textitelbow que resultou em 3. Foram alcançados resultados satisfatórios dado o objetivo definido no projeto, além de demonstrar o agrupamento dos três grupos definidos utilizando bases treinadas de *scores* de sentimento.

O método apresentado por Folego et. al [22] classifica e identifica pinturas do artista Vincent Van Gogh empregando redes neurais e algoritmos de aprendizado de máquina. A autoria de uma pintura é um caso de extrema relevância do ponto de vista histórico-social e de valor monetário.

Especialistas em arte usam múltiplos métodos de aquisições de imagens para estudar características únicas de cada artista determinando assim sua autoria e autenticidade. Uma das contribuições desse estudo está no conjunto de imagens digitais de alta qualidade, 196.3 pixels por polegadas, das pinturas do artista Vincent Van Gogh com um total de 124 itens, onde 20% foi separada para teste.

Após o conjunto de dados ser obtido o autor define o *pipeline* do seu projeto em três fases: fragmentação, extração de características e classificação. Primeiramente dividem-se as imagens em fragmentos de 224x224 pixels, pois, como as imagens possuem diferentes tamanhos consegue-se diminuir o custo computacional sem perder o nível de detalhes pôr as mesmas estarem em alta resolução. Para a extração de características foi utilizada uma rede neural convolucional, que são redes neurais capazes de extrair padrões através de processos de filtragem e convoluções usados principalmente para imagens, previamente treinada com um massivo conjunto de dados (1.3 milhões de imagens) que tornaram a rede capaz de extrair padrões visuais complexos de imagens. Essa rede está disponível no repositório do ImageNet.

Com a rede neural convolucional conseguiu-se extrair 4096 características para cada fragmento, logo esse conjunto é apresentado como entrada para um classificador utilizando máquina de vetores de suporte. Esse mapeia os objetos em um hiperplano no qual consegue-se distinguir as respectivas classes, determinando se o fragmento pertence ao artista Vincent Van Gogh ou não.

A partir dos scores de cada fragmento define-se a classe da pintura. Conseguindo um desempenho de classificação de 92,3%, porém, com um número alto de falsos positivos dado ao alto grau de semelhança de certas pinturas do artista com os movimentos artísticos da época da sua vida.

Esse trabalho tem como contribuição a utilização de redes neurais para aplicação de extração de características de imagens complexas como as pinturas de Van Gogh e a ferramenta desenvolvida para o auxílio de especialistas da arte nos estudos das mais

diversas obras.

No caso de Alshari et al. [23], temos como principal foco de seu estudo a análise de sentimentos via o agrupamento de palavras no espaço vetorial. Mikolov et al. propuseram o modelo de representação vetorial para palavras denominado Word2Vec que basicamente transforma as palavras em valores vetoriais fundamentados em sua relação com os N-gramas. Esse método revela a relação semântica entre as palavras de um mesmo documento ou textos. Como base de dados, esse trabalho utilizou os comentários de avaliação da plataforma IMBD relacionada a filmes e séries.

O estudo propõe que inicialmente uma conversão do texto para o espaço vetorial utilizando o modelo Word2vec. Após essa etapa é realizado o agrupamento de palavras pelas operações de similaridade, com isso temos clusters oriundos da distribuição de polaridades baseado em centroide pré-determinados por um dicionário de léxico sentimental, não informado a fonte pelo autor. Os grupos formados com teor negativo sofrem uma simples transformação, para uma melhor separação da distribuição no espaço. Dado que os vetores disponibilizados pelo Word2Vec são de alta complexidade o trabalho sugere o uso apenas dos centroides, definidos dada a similaridade ou a média das similaridades (mais de um termo) entre os vetores do documento, como característica principal.

Posteriormente a todos esses processos, são empregados classificadores com o intuito de definir as polaridades de cada comentário dado as novas características definidas em cada grupo formado. Foi obtida nesse estudo uma desempenho de acerto de 93,80% utilizando regressão logística como o classificador. Apesar de o resultado de alta rendimento o estudo não deixa claro de como a redução da complexidade através de seu método pode realmente representar o sentimento de determinado corpo textual. Entretanto, os resultados alcançados e a contribuição em questões de análise de sentimentos com texto no espaço vetorial indicam um amplo campo ainda a ser estudo e desenvolvido em trabalhos futuros.

Identificação de artistas na área de belas-artes é um grande desafio executado por especialistas da arte com anos de experiência e estudo. Algoritmos de aprendizado de máquina e redes neurais são ferramentas interessantes para o reconhecimento de padrões e classificação dos mesmos para determinar a autoria de uma obra artística. Em [24] exemplifica-se o treinamento de modelos simples desde redes neurais convolucionais desenvolvidas para o problema assim como empregando a transferência de conhecimento a partir de redes previamente treinadas.

Esse estudo utilizou um conjunto de dados que contém cerca de 300 pinturas para 57 artistas marcantes na história da arte, sendo dividido 80% para treinamento, 10% para teste e 10% validação. Para às duas redes estudas o pré-processamento aplicado nesses

dados foi o mesmo, reduzindo as imagens em fragmentos de 224x224 pixels e randomicamente rotacionando-as para evitar um possível textitoverffting.

Foram utilizadas no treinamento três redes convolucionais, utilizando classificador *softmax* com entropia cruzada, logo essas redes constantemente tentam maximizar o score de acerto para cada artista. A primeira rede a ser utilizada foi uma com arquitetura tradicional convolucional, portanto, é realizada uma série de reduções na entrada para diminuir a complexidade das entradas e agregar suas principais características.

A rede com arquitetura ResNet-18 foi empregada em dois cenários: desenvolvida do zero e com transferência de conhecimento; a ResNet usa blocos residuais para garantir que os gradientes sejam propagados para todas as camadas da rede inclusive as camadas mais profundas, permitindo um aprendizado mais sólido de determinada característica.

A transferência de conhecimento é uma técnica no qual temos uma rede treinada previamente (ImgNet), ou seja, seus pesos e bias já foram determinados, portanto, podemos apenas modificar a última camada para ajustar para um dado problema, no caso a identificação de artistas.

Nesse estudo foi utilizada a biblioteca PyTorch além de que todos os modelos seguiram, no treinamento, o uso da regra de atualização da rede denominada Adam. A métrica utilizada foi F1-Score, reportando assim a relação entre falsos positivos, falsos negativos e totalmente corretos.

Esse estudo mostrou um resultado quantitativo de 71,0% utilizando rede convolucional clássica, 73,3% para ResNet desenvolvida do zero e 89,8% com a transferência de conhecimento. Os resultados obtidos são a estado da arte e o estudo demonstrou o poder das redes convolucionais de extraírem padrões e características de imagens de obras das belas-artes, além de uma análise qualitativa junto com mapas de saliência comprovando a eficácia dessas redes para o reconhecimento de padrões de uma determinada classe, no caso dos 57 artistas estudados.

Em [25] são exemplificadas as diversas abordagens frente a extração de emoções em texto assim como seus respectivos métodos, visto que o tema nos últimos tem ganhando cada vez mais importância na área de processamento de linguagem natural. A detecção de emoções em documentos é baseada em métodos de classificação que, em conjunto com os conceitos de Linguagem Natural e Aprendizado de Máquina, são capazes de determinar o sentimento e a subjetividade presente naquele documento. Sentimentos são determinados pela psicologia social em W. Gerrod Parrot e Ekman, onde os mesmos sugeriram cinco classes de emoções: amor, alegria, raiva, tristeza, medo e surpresa; logo, define-se como estados mentais acompanhados de mudanças psicológicas entre os próprios.

No campo da NPL, a classificação de sentimentos pode ser fundamentada em um *score* dado a cada palavra dentro de um corpo de texto agregando cada token em seu resultado.

Também são utilizados métodos nos quais se analisam as relações entre os sujeitos de cada verbo com seu significado e ainda modelos de aprendizado de máquina antecipadamente treinados que junto a características linguísticas determinam a polaridade de um documento.

O trabalho ao mesmo tempo, propõe três métodos de detecção de sentimentos. O primeiro o enfoque é na detecção baseada em palavras chaves do texto, ou seja, um conjunto de palavras com importância significativa em um dado documento. Com isso assimilasse seus sinônimos e antônimos na WordNet — dicionário léxico com a associação das respectivas polaridades — para prever sua orientação semântica e sua polaridade.

O segundo procedimento consiste na identificação de sentimentos via um modelo empregando o algoritmo de máquina de suporte vetorial treinado com uma grande quantidade de dados previamente rotulados. Por último é apresentada a metodologia híbrida em que é utilizada a polaridade dos dicionários de palavras WordNet para melhorar a desempenho do classificador, ou seja, adiciona-se a polaridade das palavras como uma das entradas do modelo.

A pesquisa apresentada deixa claro o espaço para outros estudos relacionados ao assunto como uso de análise de emoções para auxiliar em sistemas de recomendação assim como em projeto de inteligência governamental e de negócios. Portanto, o autor conclui os métodos disponíveis atualmente, funcionam bem quando aplicados para um escopo controlado, entretanto, pela ambiguidade da linguagem e vocabulários específicos em cada tema existe ainda muita a se melhorar em desempenho.

O trabalho [26] propõe um método para avaliar a complexidade de uma certa imagem baseada em redes neurais, tendo como parâmetros de entrada a textura, informações de borda e área de significância. A complexidade de imagens é uma importante vertente de pesquisa na área de processamento de imagens e reconhecimento de padrões, uma vez que é usada para determinar o grau de dificuldade para extrair ou reconhecer um objeto, ou alvo em uma certa imagem.

Em trabalhos de referência e estudados o autor reitera que foram utilizadas características de contraste, intensidade de bordas e textura e ainda possui uma análise subjetiva por parte de avaliações com pessoas e ajustes dos pesos e índices de forma manual. Dessa forma, o estudo propõe o uso de redes neurais para não depender de fatores ou erros humanos além de um resultado com uma robustez matemática aperfeiçoada.

Como entrada da rede neural foram escolhidas três características principais como citadas anteriormente: textura, bordas e área significante. Para a textura foi utilizada a matriz de coocorrência em escala de cinza que disponibiliza três parâmetros para cada imagem: o contraste que descreve a clareza da textura, a correlação que é usada como grau de similaridade de padrões na imagem, a energia que descreve a distribuição uniforma da

imagem em escala de cinza. As bordas foram descritas pelo algoritmo de Sobel, tendo como resultado a taxa de quantidade de bordas que reflete a complexidade da imagem. Na área de significância utilizou-se o coeficiente da transformada discreta de cosseno.

Definida as entradas, escolheu-se a rede neural BP que pode ser usada para modelar processos não lineares onde não é preciso determinar o modelo dos dados. A rede é da categoria de aprendizado supervisionado logo com o conjunto de treinamento e utilizando *back-propagation* a rede ajusta seus pesos e bias até que o quadrado de seus erros se reduza ao máximo. Ao final, a rede possui na camada de saída uma função *sigmod* que estabelece o resultado, a mesma foi treinada com 200 imagens com diferentes complexidades.

Em comparação com os métodos de referência e uma análise de complexidade subjetiva realizada com pessoas, os resultados obtidos seguiram uma tendência similar e até de forma mais robusta, provando que a utilização de redes neurais para determinar os pesos e índices de complexidade são um método efetivo.

Capítulo 5

Materiais e Desenvolvimento

O trabalho proposto visa correlacionar as emoções e sentimentos contidos nas correspondências de Van Gogh, considerando a complexidade de suas obras visuais e objetivando auxiliar a análise feita por especialistas e curadores da arte sobre a biografia do artista. A solução expande-se em duas frentes de estudo: suas cartas e suas pinturas. As cartas são as principais fontes de sua biografia e possuem vasto conteúdo sobre sua vida e descrições sobre sua obra artística.

Consequentemente, adotamos como abordagem o uso dos algoritmos e técnicas de mineração de textos e processamento de linguagem natural, para a obtenção de características contidas nas cartas sobre a personalidade do artista em diferentes passagens de sua trajetória.

O artista passou por sérios distúrbios psicológicos, atualmente caracterizados como, por exemplo, transtornos bipolares e depressão. De forma simplista, esse estado de saúde mental possui forte influência com a parte criativa de suas obras elaboradas em sua carreira artística.

Na época, não ficou claro quais seriam realmente os transtornos psicológicos de Van Gogh, portanto, sempre foi taxado com o rótulo de "louco". Com esse cenário, propõe-se uma análise de emoções e sentimentos de suas cartas, relacionando-as com as definições psicológicas predominantemente em cada etapa de sua vida.

Em um breve estudo do domínio problema, foram constatados dois problemas: número limitado de dados (cerca de 1000 documentos) e nenhum dado rotulado previamente. Dessa forma, foi estabelecido que os algoritmos e técnicas aplicadas deveriam suportar um número restrito de exemplos, além de conseguirem trabalhar com dados de maneira não supervisionada, uma vez que os mesmos nunca foram rotulados para essa atividade.

Com o problema bem definido e empregando o *pipeline* de processamento de linguagem natural, dividiu-se o método proposto nos seguintes segmentos: aquisição dos dados, pré-

processamento, representação do texto e aplicação de algoritmos de agrupamento com o intuito de classificar as emoções e sentimentos embutidos no texto.

Como não foi encontrado uma base de dados no qual poderíamos ter acessos aos textos indexados e estruturados de forma a facilitar a sua manipulação, foi proposto a obtenção dos mesmos por meio do sítio oficial do Museu Van Gogh, visto que é a fonte mais confiável para a aquisição dos dados. A coleta ocorreu através de rotinas de *webscraping*, nas quais realizam o *download* de páginas HTML e a partir das estruturas desses arquivos é possível adquirir as informações desejadas.

Os dados obtidos foram armazenados em um diretório específico. Devido à base de dados ter sido feita via *webscraping*, é necessário realizar procedimentos de limpeza de dados para eliminar e corrigir símbolos não textuais que foram extraídos de forma errônea. Além disso, para melhorar o desempenho em futuros algoritmos de processamento de texto foram aplicados a remoção de *stopwords* e eliminação da pontuação.

Com os dados provenientes das tarefas anteriores, separamos o tratamento dos dados em duas etapas: a parte descritiva e o reconhecimento de padrões.

O entendimento do problema proposto deve-se muita das vezes a descrição dos dados obtidos. Por isso foi empregado, uma série de rotinas e algoritmos para obter características e estatísticas gerais, ou seja, que englobam todos os respectivos dados. Também foi feito esse processo para cada documento de forma separada.

Já na parte preditiva, tivemos dois objetivos centrais: a identificação de emoções em cada documento e da correlação entre o vocabulário e a biografia de Van Gogh. Quanto a análise de sentimentos, subdividimos em dois mecanismos: extração de sentimentos (positivo, negativo e neutro) utilizando *SentiNetWord* e a identificação de emoções através de um dicionário léxico das oito principais emoções definidas na teoria de Plutchik [2]: raiva, felicidade, nojo, tristeza, medo, surpresa, antecipação e confiança.

O processo de análise de sentimentos com *SentiNetWord* é realizado consultando um dicionário no qual cada palavra possui uma pontuação definida para cada um dos polos positivos e negativo, assim como o neutro. Com a pontuação de cada palavra, calcula-se média ponderada, sendo os maiores pesos equivalentes aos adjetivos e advérbios, pois, essas estruturas são as que representam de forma mais objetiva os sentimentos em um texto.

A outra análise é baseada na ocorrência de palavras que representam de alguma forma a expressão de uma das oito emoções escolhidas. Logo, os dicionários possuem 8 seções, que possuem palavras em cada uma delas, palavras fortemente relacionados a determinada emoção. Portanto, é computado a frequência das palavras em cada um dos textos de entrada.

Com a intenção de correlacionar as emoções e os sentimentos contidos em cada documento, foram utilizados os algoritmos de agrupamento visto que os nossos dados não possuem um rótulo pré-definido, escolheu-se o algoritmo *K-means* amplamente utilizado na literatura.

Além da parte de análise textual, foi realizada um *pipeline* semelhante com a obra artística de Van Gogh com a finalidade de avaliar a complexidade de cada uma de suas pinturas e desenhos ao longo de sua carreira. Dessa forma, basicamente repitimos os procedimentos de aquisição de dados e extração de características, voltado para imagens, além da análise da complexidade.

Na aquisição de dados, utilizamos novamente o *webscraping* para obter os dados de forma indexada e organizada. Em seguida, os algoritmos de extração de características foram aplicados gerando entradas para as análises de complexidade e cores.

A Figura 5.1 representa a solução proposta de forma gráfica para facilitar o entendimento.

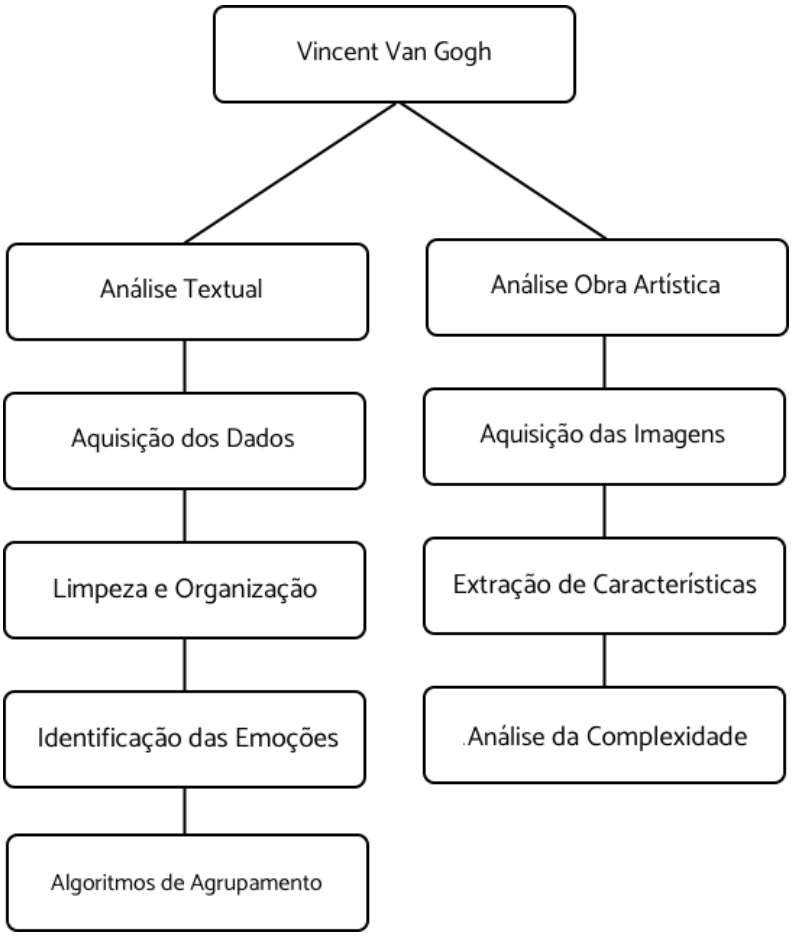


Figura 5.1: Fluxograma da solução proposta com cada uma das etapas

5.1 Análise Textual

5.1.1 Base de Dados

Para a aquisição das informações textuais, foram utilizadas as técnicas de *webscraping* com o intuito de coletar de forma automática os textos das cartas escritas e enviadas ao artista. Essa coleta foi realizada no *website* oficial do Museu Van Gogh, que possui o maior acervo sobre sua vida. Esse possui um total de 928 cartas, sendo essas escritas pelo pintor para o seu irmão Theo, os pintores pauis Gauguin e Emile Bernard, entre outros amigos e familiares.

Como citado acima, foi feita a coleta de dados na web via o *download* de documentos *HTML* transformando-os as informações textuais em estruturas de dados que serão utilizadas como entradas em futuras análises.

O algoritmo desenvolvido pode ser dividido em duas etapas:

- **Indexar cada carta com seu link correspondente**: foi baixado o HTML de http://vangoghletters.org/vg/letters.html e coletado junto a tag de listagem os títulos e links de cada carta.
- **Efetuar o download do conteúdo junto ao link único**: varreu-se cada link para extrair os textos eliminando possíveis tags HTMLs e símbolos insignificantes, geralmente contidos nesses tipos de arquivo.

Ao final de todas as etapas, adquirimos um conjunto de dados indexados pelo título e o conteúdo textual, igualmente ao apresentado no sítio do Museu Van Gogh. Esse processo pode ser exemplificado pelo fluxograma da Figura 5.2.

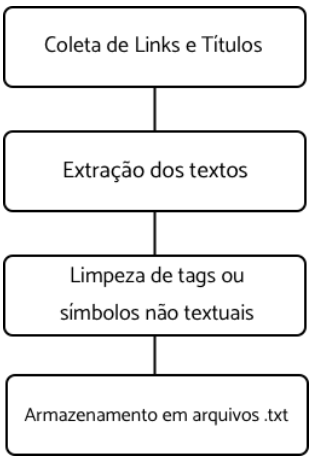


Figura 5.2: Fluxograma de Aquisição das Cartas no sítio do Museu Van Gogh

Com a etapa de aquisição concluída, constatou-se que essa base serve como uma contribuição para futuras pesquisas na área de processamento de linguagem natural sobre Vincent Van Gogh, uma vez que os dados pertinentes em forma de texto foram organizados em diretórios e estrutura de dados, facilitando assim seu acesso e manipulação.

5.1.2 Pré-Processamento e Limpeza de Dados

O pré-processamento realizado no aglomerado de informações faz-se necessário para um melhor desempenho dos algoritmos de aprendizagem de máquina, uma vez que ruídos e erros de aquisição são comuns em um número grande de exemplares.

Quando tratamos de texto devemos realizar uma etapa extra, além dos processos já conhecidos em mineração de dados e IA. O pré-processamento consiste em reduzir o número de palavras no texto para apenas aqueles símbolos que agreguem valor semântico e sintático para uma análise textual.

O resultado dessas etapas ocorre seguindo os seguintes procedimentos:

- **Remoção das *Stopwords***: definiu-se um conjunto de palavras que possuem uma frequência elevada e, ao mesmo tempo, não tem significado para texto como, por exemplo, os artigos da língua inglesa “the” e “a”; e assim removidas dos textos de entrada.
- **Remoção de Símbolos Indesejados**: como o texto foi extraído de páginas web, os mesmos possuem símbolos e caracteres que não contribuem para o conteúdo dos textos, os mesmos também são removidos.
- **Remoção da Pontuação**: a pontuação foi removida dos textos para uma melhor análise em determinados casos.

Como exemplo, temos um trecho de uma das cartas e o resultado da limpeza sendo expressado tanto em formal textual como na visualização da nuvem de palavras (Figura 5.3 e Figura 5.4).

5.1.3 Organização de Dados

Para melhorar o processo de análise, é necessária uma organização dos elementos adquiridos nos procedimentos anteriores, além de realizar uma decomposição descritiva de cada documento e de seus metadados.

Foram escolhidos de forma empírica as características relevantes para exploração e aplicação de algoritmos não supervisionados com o intuito de encontrar padrões entre os documentos e as pinturas de Van Gogh.

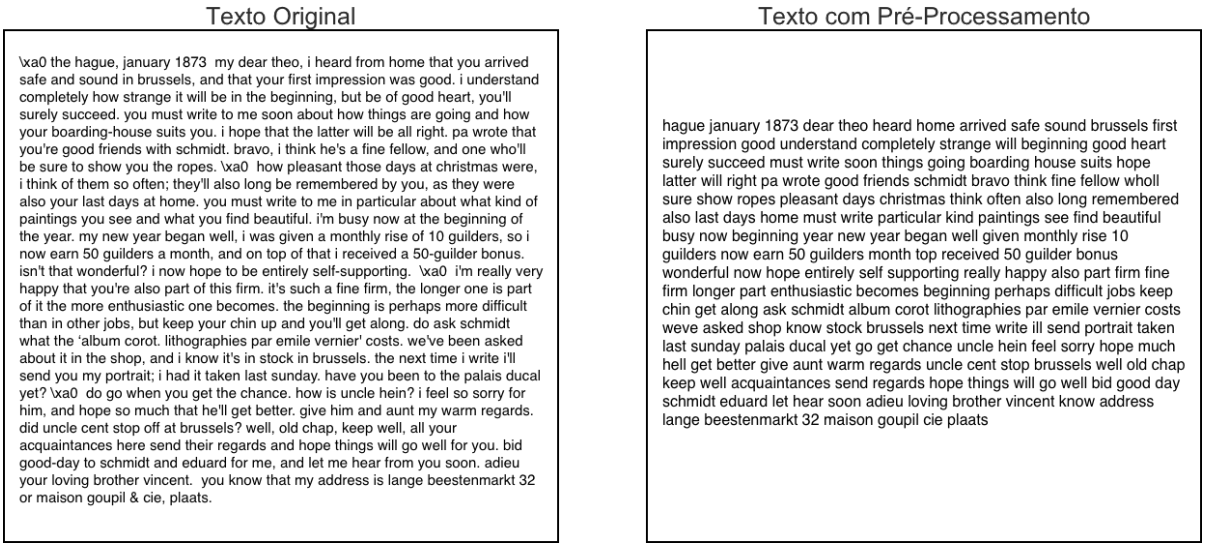
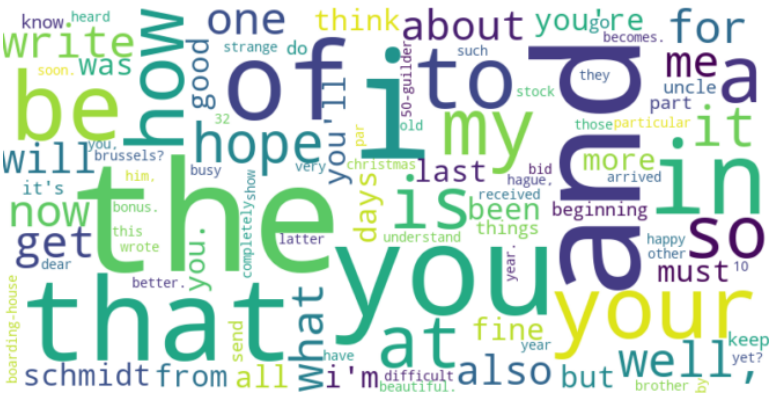
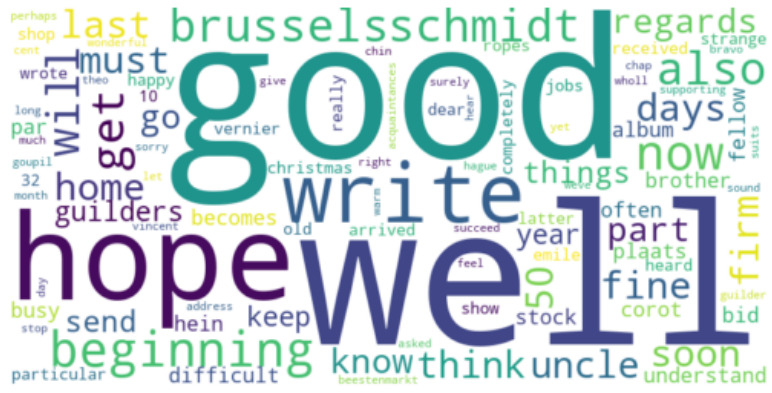


Figura 5.3: Comparação de textos antes e depois do pré-processamento



(a)



(b)

Figura 5.4: (a) WordCloud Texto de Entrada (b) WordCloud Texto após o pré-processamento

index	numberLetter	letterName	dateLetter	numWords	FromToVanGogh	mostFreqWords
0.0	112	To Theo Van Gogh	23 April 1877	821	True	will, hope, got
1.0	607	To Theo Van Gogh	10 May 1888	556	True	need, put, like
2.0	655	To Emile Bernard	5 August 1888	1598	True	like, well, hard
3.0	434	To Theo Van Gogh	9 March 1884	984	True	millet, think, something
4.0	190	To Anthon Van Rappard	23 November 1881	1450	True	love, well, man
5.0	016	To Caroline Van Stockum- Haanebeek	20 November 1873	125	True	good, hope, wish

Figura 5.5: Exemplo do Conjunto de Dados após limpeza e mineração

As informações disponíveis estão discriminadas para cada artefato em: índice, título, autor e data da escrita, o número de palavras, as palavras e ngrams mais frequentes.

Foi seguido o *pipeline* padrão de mineração de dados, no qual deve-se realizar a visualização dos dados e características para melhor entendimento domínio do problema. Essa questão se deu no uso de nuvens de palavras, apuradas junto a rotinas desenvolvidas ao longo do trabalho. (Figura 5.5).

Ao final, forma-se uma série de estatísticas onde é possível um julgamento descritivo de todos os dados obtidos, como o número de cartas por ano, palavras mais frequentes em relação a todo o aglomerado e assim demonstrar a biografia do artista em prol dos respectivos corpus existentes.

5.1.4 Análise de Sentimentos

Para identificar os sentimentos em cada um dos textos e de forma geral, empregaram-se dois procedimentos: utilizando o dicionário léxico *SentiNetWord* e uma nova metodologia proposta por este trabalho denominado dicionário de emoções.

Análise Sentinetword

Realizar a classificação de textos em positivo e negativo com um conjunto de dados abstrato, ou seja, sem possuir dados de treinamento com um rótulo definido para realizar o *pipeline* padrão de classificação por meio das etapas: treinamento, validação, otimização e testes com um novo conjunto de dados exigem outros tipos de abordagem. A escolhida para esse trabalho foi a análise léxica no qual se verifica o *score* definido de cada palavra do texto entre os dois polos.

O *SentiWordNet* 3.0 reúne o conjunto de palavras disponíveis no *WordNet*, o maior dicionário léxico da língua inglesa, sendo que cada palavra possui um *score*, tendo como entrada a sua função sintática, para as polaridades: positivo, negativo e neutro. Dessa forma, foi realizado o *pipeline* de análise das emoções presente na Figura 5.6

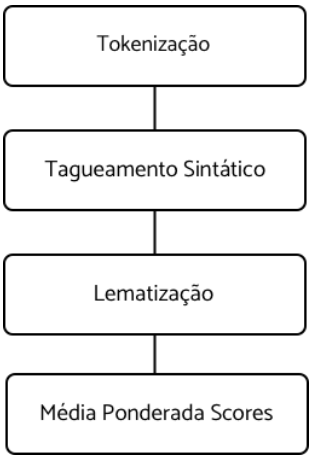


Figura 5.6: Pipeline para análise sentimentos utilizando SentiWordNet

O primeiro procedimento a ser a realizado é a tokenização, esse processo utilizou a heurística da pontuação e espaços do texto como base para realizar o fracionamento de cada palavra. Com esse novo conjunto de dados formatado, ocorre a identificação da classe gramatical que consiste em determinar afunção sintática de cada palavra, como, por exemplo, na Figura 5.7. Dessa forma conseguimos retirar, em parte, a ambiguidade dos textos.

Van Gogh loves to paint at afternoon.

Figura 5.7: Exemplo de tagueamento da frase: "Van Gogh loves to paint at afternoon.". Em vermelho os verbos, em cinza os substantivos, em amarelo as preposições e preto estruturas axiliares.

Sabendo a função sintática de cada palavra do texto, realiza-se a lematização, que em termos linguísticos significa obter puramente o lema, o sentindo da palavra, não considerando por exemplo: o tempo verbal, gênero, número e até mesmo sufixos e prefixos. Essa é a forma clássica de eliminar ao máximo a ambiguidade em uma análise léxica.

Após todas essas etapas, consulta-se o dicionário de *scores* no *SentiWordNet* de cada palavra do texto, realizando uma média ponderada para cada dos estados positivo e negativo. Contudo, resolvemos retirar o neutro por não contribuir em nossas análises. Ao final atribuímos maior peso aos adjetivos e advérbios, uma vez que essas estruturas representam puramente o sentimento de um texto.

Ao final, temos um *score* total de cada texto em uma escala entre 0 e 1, no qual 1 indica que o positivo, assim como o zero, o negativo.

Análise Dicionário de Emoções

Dividir os sentimentos contidos em um texto apenas em positivo e negativo é de certa forma raso, pois, podemos considerar que o conteúdo por ser um meio de comunicação complexo possui inúmeros significados associados assim como diversas emoções envolvidas.

Baseando-se na teoria psicológica das emoções procurou-se uma investigação das ocorrências de palavras que representam de alguma maneira, as emoções das seguintes categorias: raiva, nojo, tristeza, alegria, medo e surpresa.

Foi proposto um conjunto de palavras, retirados de bases robustas (dicionários da língua inglesa), que de certa forma são próximas para as categorias escolhidas. Com o dicionário proposto e utilizando como exemplo o modelo *bag of words*, computamos a frequência desses sentimentos em cada um das entradas.

Exemplifica-se essas estruturas a seguir. Os exemplos estão em Língua Inglesa visto que os textos foram disponibilizados dessa forma:

- **Alegria:** contentment, pleasure, contentedness, satisfaction, cheerfulness, cheeriness, merriment, merriness, gaiety, joy, joyfulness, joyousness, joviality, jollity, jolliness, glee, blitheness, carefreeness, gladness, delight, good spirits, high spirits, lightheartedness, good cheer, well-being, enjoyment, felicity, exuberance, exhilaration, elation, ecstasy, delirium, jubilation, rapture.
- **Tristeza:** sad, unhappiness, sorrow, dejection, regret, depression, misery, cheerlessness, downheartedness, despondency, despair, desolation, wretchedness, glumness, gloom, gloominess, dolefulness, melancholy, low spirits, mournfulness, woe, brokenheartedness, heartache, grief, unfortunate, regrettable, sorry, wretched, deplorable, lamentable, pitiful, pitiable, pathetic, shameful, disgraceful, tragic, unhappy, awful, sorrowful, miserable.
- **Raiva:** annoy, vexation, exasperation, crossness, irritation, irritability, indignation, pique, displeasure, resentment, enrage, fury, wrath, outrage, temper, road rage, air rage, irascibility, ill temper, dyspepsia, spleen, ill.
- **Medo:** terror, fright, fearfulness, horror, alarm, panic, agitation, trepidation, dread, consternation, dismay, distress, anxiety, worry, angst, unease, uneasiness, apprehension, apprehensiveness, nervousness, nerves, timidity, disquiet, disquietude, discomposure, unrest, perturbation, foreboding, misgiving, doubt, suspicion, the creeps, the willies, the heebie-jeebies, the shakes.

- **Surpresa:** astonish, amaze, nonplus, startle, astound, stun, flabbergast, stagger, shock, stupefy, leave open-mouthed, dumbfound, daze, benumb, confound, take aback, jolt, shake up, bowl over, knock for six, floor, strike dumb, take by surprise, catch unawares, catch off guard.
- **Nojo:** sicken, outrage, offend, sicken, outrage, offend, revolt, put off, repel, nauseate, turn your stomach, fill with loathing, cause aversion, loathing, revulsion, hatred, dislike, nausea, distaste, aversion, antipathy, abomination, repulsion, abhorrence, repugnance, odium, detestation.
- **Antecipação:** apprehension, hope, joy, prospect, contemplation, expectancy, foresight, foretaste, impatience, outlook, preconception, premonition, preoccupation, prescience, presentiment, promise, trust, awaiting, high hopes, looking forward, expectation, forethought, readiness for.
- **Confiança:** confidence, expectation, faith, hope, assurance, certainty, certitude, conviction, credence, credit, dependence, positiveness, reliance, stock, store, sureness, entrustment, gospel truth, acceptance, assuredness, surety, dependence (also dependance), cartel, combination, combine, syndicate, chain.

5.1.5 Associação das Emoções

Para o problema em questão, utilizaram-se algoritmos de aprendizagem de máquina não supervisionados e modelos que representam as palavras em vetores multidimensionais.

O conjunto de dados construídos ao longo das últimas seções não possui um rótulo, logo utilizou-se o algoritmo de agrupamento para validar os grupos de emoções, assuntos das respectivas cartas, estágios da vida de Van Gogh, suas pinturas e a sua biografia.

Por isso, foi utilizado o algoritmo *K-Means* para definir os grupos de acordo com a distância entre os exemplos da base de dados. Visto o interesse em encontrar diversos cenários, os dados foram serializados pelo *score* dos sentimentos, utilizando*SentiWordNet* e a frequência do dicionário léxico de emoções.

Para cada conjunto aplicou-se a normalização dos dados, já que o algoritmo utilizado é baseado na distância entre os elementos. Para escolha dos centroides, foi aplicada a curva de *elbow*, que auxilia na definição do k para o nosso conjunto de dados. Adotou-se $k = 4$, visto os testes alcançaram resultados satisfatórios.

Ainda foram empregados os modelos que representam as palavras e documentos como vetores multidimensionais. De forma exploratória e de validação do método proposto, ocorreu o treinamento dos algoritmos Word2Vec (utilizando todas as cartas como um único texto).

```
indexes, metrics = model.cosine('family')
```

```
model.vocab[indexes]
```

```
array(['wife', 'father', 'son', 'feelings', 'soul', 'illness', 'presence',
       'voice', 'countenance', 'eyes'], dtype='<U78')
```

(a)

```
indexes, metrics = model.analogy(pos=['expression','painting'], neg=[],n=10)
```

```
model.generate_response(indexes, metrics).tolist()
```

```
[('sentiment', 0.7003059081519951),
 ('print', 0.6720221039209211),
 ('subject', 0.6681731382321997),
 ('type', 0.6663359416472168),
 ('style', 0.6442394394539144),
 ('basis', 0.637667756997643),
 ('structure', 0.6355628652225674),
 ('figure', 0.6355001233554363),
 ('quality', 0.6330166489367358),
 ('bunch', 0.6325462415789329)]
```

(b)

Figura 5.8: Exemplo de manipulações com Word2Vec: (a) Vizinhança de Palavras (b) Analogia a partir da soma de duas palavras (vetores)

Para o Word2Vec, verificou-se a presença de algumas palavras próximas que se conectam com diversas vertentes da sua vida. Por exemplo, a proximidade das palavras referentes a sentimentos e até mesmo momentos da sua vida como exemplificado na Figura 5.8.

5.2 Análise da Obra Artística Visual

5.2.1 Base de Dados

Como não há uma base de dados disponível de suas pinturas e desenhos, utilizou-se da técnica de *webscraping*, descrita anteriormente, para extrair de forma automatizada todos essas imagens. O sítio utilizado foi o WikiArt, que consiste em uma enciclopédia digital onde são disponíveis os trabalhos artísticos de grandes personalidades e das mais variadas épocas.

Para essa coleta de dados, foi feito algo semelhante ao descrito na aquisição de dados textuais:

- **Downalod da página principal:** foi baixado o HTML de https://www.wikiart.org /en/vincent-van-gogh/all-works/text-list, assim armazenou-se o link de cada obra do artista disponível.

- **Indexados cada imagem pelo seu nome e ano:** com o intuito de ter disponíveis informações básicas de cada pintura e de seus desenhos foram obtidos o nome, ano e gênero(pintura ou desenho) de cada item.
- **Efetuar o *download*das imagens:** foi realizado o *download* de cada imagem na melhor resolução possível e armazenadas em diretório específico.

Após todo o procedimento, temos uma base de dados indexadas por nome, ano de finalização da obra e seu tipo podendo ser pintura ou desenho. Como foi disponibilizado de forma pública com o intuito de contribuir para futuros trabalhos relacionados a obra de Van Gogh visto que estão indexados e contém pinturas e desenhos, diferentes de outras bases de dados.

5.2.2 Extração de Características

Com a disponibilização de todas as imagens de sua carreira astísica, realizamos os procedimentos de extrair de cada imagem suas características com o intuito de avaliar a complexidade das mesmas.

Utilizamos três meios de obter características: histograma, paleta de cores, GLCM e bordas.

O histograma de uma imagem colorida descreve a relação com as cores de cada imagem e a ocorrência das mesmas. Outra característica relacionada as cores são as paletas de cores onde se realiza a quantização das imagens com o objetivo de obter as n cores principais da mesma.

(a) (b)

Figura 5.9: (a) Imagem Colorida (b) Paleta de Cores da Imagem Colorida com 8 cores

Através da criação da Matriz de Co-ocorrência de Níveis de Cinza(GLCM) de cada canal de cor do sistema RGB, conseguimos características referentes a textura de cada imagem existente como: energia, correlação, contraste e homogeneidade.

Além disso, foram colhidas informações sobre as bordas de cada imagem utilizando o algoritmo de Canny.

5.2.3 Extração da Paleta Cores

Em especial para as cores, vale destacar alguns procedimentos realizados para a extração das diversas paletas de cores de uma imagem, da mesma maneira ressaltar a metodologia utilizada para avaliar a esparsidade entre as cores determinadas como chaves para cada uma das imagens.

Em primeiro momento, foi realizado a clusterização de cada quadro do artista utilizando o algoritmo *K-means*. Para a escolha do número de *clusters*, utilizou-se das referências de quantização da representação de 8 bits(245 tons), logo foram geradas paletas de 8, 16, 32, 64 e 128 cores.

Portanto, para cada quadro foram extraídas cores representadas por centroide no espaço RGB. Para cada um desses centroides encontrados foi também proposto a proporção de pixels de uma certa cor na imagem, para que as cores com maior número de pixels tenham centroides garantidos.

Com essa etapa da extração dos centroides, temos como resultado a paleta de cores dentro de um determinado quadro. Que para métricas anuais, foi realizada ma nova iteração do *K-means* com o intuito de agrupar as principais cores dado um conjunto de imagens.

Além dessa extração, com o auxílio dos centroides foi calculada como métrica de esparsidade e uso de cores a distância entre os mesmos, de forma a representar de forma vetorial a variância de cores utilizadas pelo artista, tendo como origem a cor mais escura de sua paleta.

Com esses atributos extraídos foram realizadas análises a serem discutidas no capítulo seguinte, referente as cores utilizadas pelo artista e sua variância ao longo dos anos.

5.2.4 Complexidade

A complexidade, apesar de ser um parâmetro subjetivo ao olhar humano, é reproduzido através de metodologias computacionais. O método utilizado para medir a complexidade de cada imagem usa características de textura e informações sobre as bordas.

As características de texturas são disponibilizadas pela Matriz de Co-ocorrência de Níveis de Cinza, sendo elas: energia, correlação, contraste e homogeneidade; e as informações de bordas são a razão de pixels considerados bordas pelo algoritmo *Canny* e o total de pixels de cada imagem. Ao final para chegarmos a uma grandeza de complexidade

temos:

$$C = w_1 * Energia + w_2 * Homogeniedade + w_3 * Correlação + w_4 * Contraste + w_5 * Bordas, \tag{5.1}$$

sendo w_n os pesos utilizados para cada parâmetro definido empiricamente ao longo do experimento. Com o C estabelecido é possível comparar cada imagem ao seu grau de complexidade.

Capítulo 6

Experimentos e Análises

Para analisar as correlações entre obra e vida do pintor Vincent Van Gogh, foram aplicadas técnicas de mineração de texto, análise de sentimentos e extração de características de seus quadros e desenhos.

Os resultados foram divididos em escopos para facilitar a visualização e possíveis análises subjetivas visto a quantidade de dados utilizadas. Foram submetidas para a abordagem textual, 928 cartas enviadas a maioria delas para o seu irmão Theo Van Gogh. Atualmente, sua biografia é basicamente reconstruída através dessa fonte, consequentemente encontrou-se similaridade entre os resultados obtidos a partir da metodologia com os trabalhos já desenvolvidas sobre sua vida pessoal e seu processo criativo.

Em um primeiro momento, será apresentado os resultados globais considerando todas as cartas como uma só, expondo algumas características e *insights* sobre o conjunto de dados em relação à trajetória do artista.

Visualizaremos também, os resultados obtidos a partir do treinamento dos modelos Word2vec.

Quanto as pinturas e desenhos de Van Gogh será apresentado a evolução da complexidade das mesmas ao longo de sua carreira artística, a partir das características de textura, cores e boras.

6.1 Análise Textual

Seguindo a metodologia apresentada na última seção, primeiramente foi feita a aquisição dos dados junto ao sítio do Museu Vincent Van Gogh, provendo as seguintes quantidades de cartas e quadros/desenhos vistos na Tabela 6.1.

Em seguida iniciou-se o *pipeline* de processamento textual, primeiramente com a limpeza dos dados. Após as primeiras iterações de remoção de *stopwords* observou-se uma frase entre as palavras mais frequentes usando o modelo *bag of words*, apresentado abaixo:

Tabela 6.1: Total de dados obtidos na etapa de aquisição de dados.

Cartas	Quadros/Desenhos
928	1.564

"I'm would like see good things well"

Ao observar brevemente sua biografia, podemos visualizar total correlação com essa frase, pois, a mesma representa empiricamente os trechos complicados vividos pelo artista em busca de seu reconhecimento artístico.

Seguindo com o refinamento do conjunto de *stopwords* a serem removidos, ao final obtivemos as seguintes 20 mais frequentes palavras de todo o corpo textual demonstrado na nuvem de palavras da Figura 6.1.

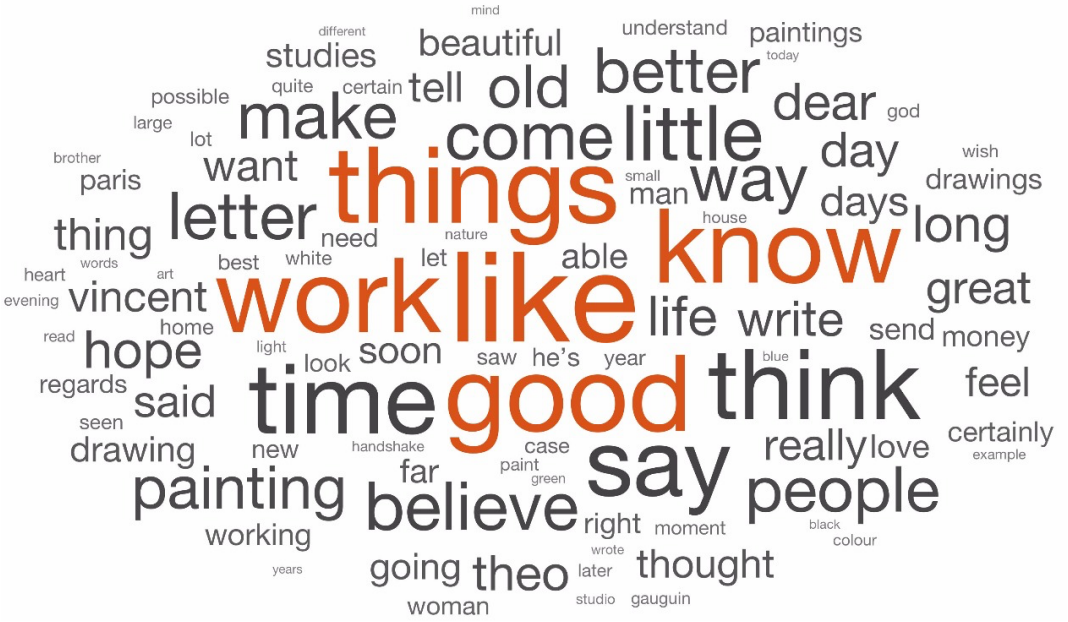


Figura 6.1: Nuvem de palavras do corpo textual de todas as cartas

Entre essas demonstradas na Figura 6.1, podemos observar palavras que em seu contexto são de extrema importância como: pinturas, trabalho, dinheiro, pessoas, desenhos, desejos, beleza, melhorar, melhor, visto, momento e acreditar. Algumas inclusive são consideradas antônimos analisando a sua vida como pinturas e dinheiro, pois, ao longo de sua carreira artística Van Gogh não conseguiu vender efetivamente nenhum quadro além de não ser reconhecido como um grande artista em sua época.

Utilizamos as palavras chaves como entrada de similaridade no modelo word2vec e conseguimos os seguintes resultados visualizados em Tabela 6.2. As palavras próximas

aos vocábulos mais frequentes dão indicativos pertinentes da sua trajetória artística, como a presença de entidades de cidades, lugares, sentimentos, artista que o influenciaram.

Tabela 6.2: Vocabulário próximo referente palavras mais frequentes do contexto.

Palavra Chave	**Vocabulário Próximo**
like	liebermann, local, technically, lepage, uhde, interiors, heads, example, israëls, strengths
things	fact, mind, thing, understand, reason, actually, matter, bad, question, think
good	wish, best, reply, underlined, transvaal, surprises, dejected, deciding, kind, truly
know	bohème, matters, deceiving, talk, instantly, nagging, hon, unwillingness, believe, reconcile
work	make, absolutely, progress, making, sell, concede, hard, consolatory, economized, renting
think	better, things, believe, certainly, thing, want, contacts, fact, try, matter
time	needing, windfall, propose, going, hell, come, russell, tersteeg, health
say	understand, question, feel, fact, mind, wrong, future, reason, thing, intervene
ill	month, pay, hermans, spend, deduct, exhausting, year, gauguin, march
little	drying, wheatfield, loosduinen, bedroom, croquis, maries
believe	bohème, think, change, risks, consider, case, business, matters, know, remain
come	needing, time, health, tell, tersteeg, windfall, stay, going, reply, hell
way	means, mind, thing, precisely, short, point, difficult, fact, sensibly, resolve
letter	writing, thank, received, mr, news, regards, vincentdear, jo, shake, glad
make	work, absolutely, progress, sensibly, making, concede, present, hard, consolatory, sell
people	arouse, mean, far, certain, oneself, amaze, painter, speculation, art, disputes
better	want, certainly, try, positively, contacts, automatically, fact, think, consider, course

painting	eaters, paintings, theories, drawing, paint, studies, working, model, format, watercolour
old	satan, halifax, church, dockyard, bandage, walk, walked, limitation,allebé,ruipérez
life	world, love, perform, heart, bond, faith, ultimately, extinguished, felt
hope	paris, hello, soon, tell, writing, going, postal, coming, pleased, write
long	happy, coughing, enquiries, correspond, wil, likelihood, cent, theos, answer, happened
write	writing, wish, friend, letter, wishes, embraced, news, underlined, wrote, visit
day	left, house, arrived, 8, lunch, train, rémy, wednesday, oclock, note
theo	adieu, vincent, handshake, goghdear, shake, vincentdear, thanks, affectionately, thank, regards
said	pa, ploughshare, family, hurt, forbid, word, ma, let, ties, inspire
great	instinctive, says, honest, yes, colliers, sense, sincerity, unwise, refer, outlook
thing	way, mind, question, precisely, fact, point, things, sensibly, means, better
want	try, better, certainly, case, course, continue, necessary, consider, contacts, absolutely
really	sure, deal, later, absorbs, start, inconvenient, year, consult, plan, sell

Dessa forma vemos que os vocabulários utilizados em todas as cartas, tem uma proximidade com sua obra artística. De maneira exploratória, percebemos que as cartas são fontes de passagens importantes da sua vida e cotidiano, visto que Van Gogh descrevia os acontecimentos com o intuito de enviar notícias para o seu irmão Theo.

Com o intuito de identificar emoções e sentimentos, as cartas foram caracterizadas pela presença de palavras relacionadas as emoções fundamentais do círculo de Plutchik(Figura 3.4): raiva, medo, tristeza, nojo, surpresa, antecipação, confiança, e felicidade.

Os resultados ano a ano das emoções identificadas são apresentados abaixo, mostrou-se interessante a evolução em torno de sua carreira artística que começou em 1881 até sua morte em 1890.

Podemos visualizar na Figura 6.2 que as emoções de felicidade e confiança estão sempre presentes em suas cartas, por essas conterem constantemente referências de agradecimento

e amor ao seu irmão Theo. Outro fator interessante é a evolução da emoção de tristeza e medo ao longo da sua carreira artística, visto uma oscilação entre os anos.

Os três últimos anos oberava-se a presença dos sentimentos de surpresa, medo, tristeza e raiva em torno de um mesmo valor. O artista nessa época estava internado no hospital Psiquiátrico em um estado de confusão mental elevado, justificando a alternância de alguns sentimentos e a presença de outros nos 3 últimos anos. Portanto, os sentimentos identificados ao longo de nossa análise, sintetiza os períodos da vida de Van Gogh, corroborando nossa metodologia e os resultados a serem apresentados a seguir.

Após essa segunda parte, foi realizado a classificação não supervisionada utilizando o algoritmo *K-means*, com o intuito de identificar padrões entre as cartas em relação às emoções previamente detectadas. Para essa iteração utilizou-se a ocorrência de emoções em cada carta e não a frequência normalizada das mesmas.

O k escolhido para o nosso conjunto de dados foi $k = 4$ determinado pela aplicação do método de *elbow* nos mesmos.

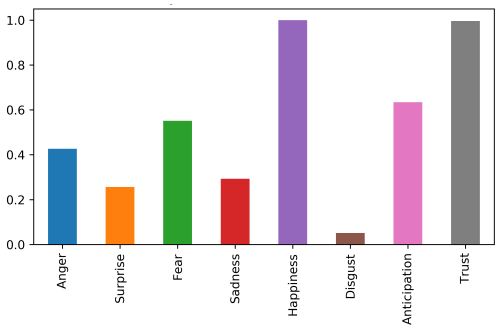
Os quatro grupos apresentaram características os quais foram associados aos *dyad*, apresentados no Capítulo de introdução teórica. Nas tabelas Tabela 6.3 e Tabela 6.4 , onde no Grupo 1 tem-se dominância das seguintes emoções: confiança, medo e antecipação; o que indica segundo a associação dos *dyad* a presença dos sentimentos de submissão, ansiedade esperança, esse grupo de emoções podem ser associados drasticamente com um comportamento ansioso por parte do artista.

No grupo dois obversa-se que a maioria de suas cartas não possuem nenhum sentimento associado e de maneira expressiva. Logo, não consideremos associações nessas cartas.

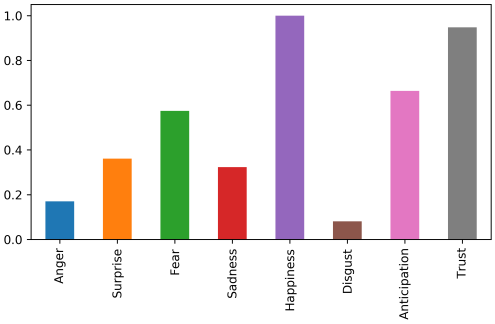
Em seguida o grupo 3 temos em todas as cartas: medo, tristeza e confiança; que pelo mapeamento do *dyad* temos os sentimentos de desespero, sentimentalismo e submissão. Esses grupo de sentimentos identificados revelam cartas contendo frustrações do artista com sua carreira, com a venda de sua arte e o reconhecimento familiar tendo ainda a sua completa dependência financeira, ao longo de alguns períodos, dependendo de seu irmão Theo.

Por último temos o grupo onde apenas felicidade e confiança demostraram grande presença que simboliza o sentimento de amor. As cartas desse grupo são muita das vezes Van Gogh trocando informações com seu irmão sobre seus momentos de felicidade ou alguma mudança positiva da sua vida artística com sua ida para o sul da França , como no início de sua parceria com Gauguim.

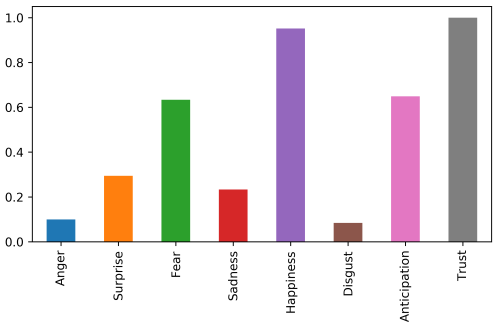
Esses dados exploratórios dão um indicativo de possíveis sentimentos ao longo de sua trajetória, que são caracterizados em sua biografia como uma confusão mental além de outros sérios problemas psicológicos não diagnosticáveis na época. Em muitas das cartas,



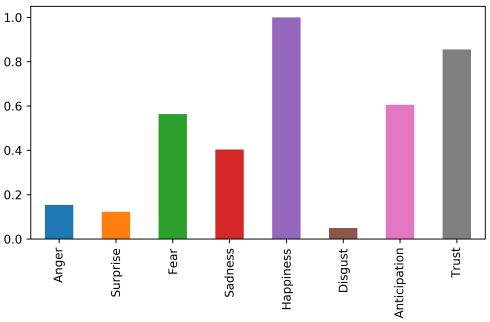
(a) 1881



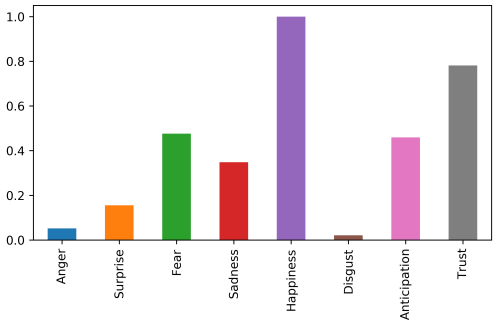
(b) 1882



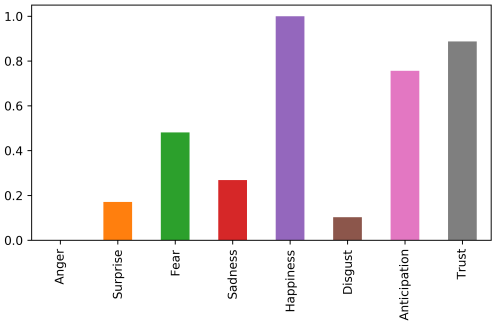
(c) 1883



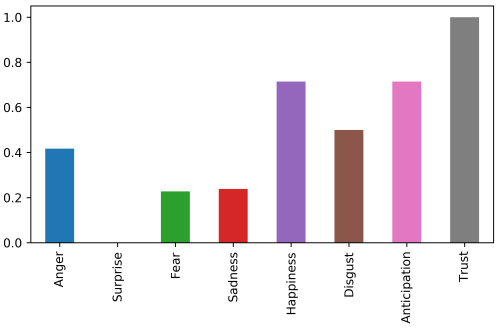
(d) 1884



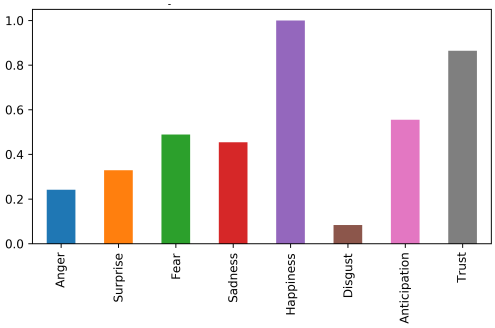
(e) 1885



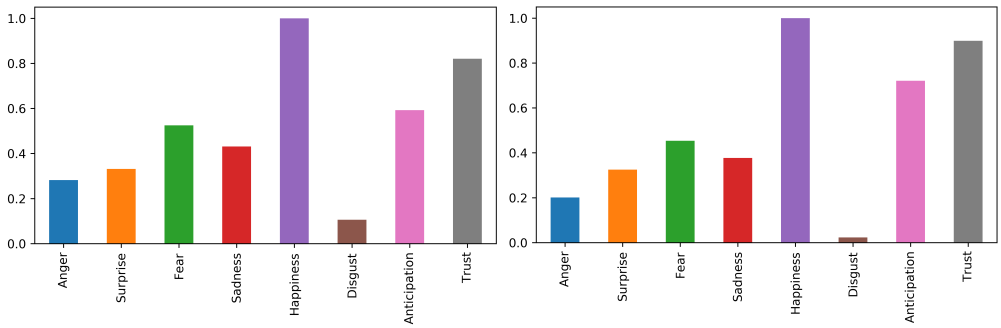
(f) 1886



(g) 1887



(h) 1888



(i) 1889 (j) 1890

Figura 6.2: São representadas ao longo de cada ano da sua carreira artísticas as seguintes emoções: felicidade, medo, tristeza, antecipação, confiança, surpresa, nojo e raiva. Os gráficos tem a referência aos anos 1881(a), 1882(b), 1883(c), 1884(d), 1885(e), 1886(f), 1887(g), 1888(h), 1889(i) e 1890(j).

Tabela 6.3: Grupos determinados pelo algortimo K-means($k = 4$)

Grupos	**Raiva**	**Surpresa**	**Medo**	**Tristeza**	**Felcidade**	**Nojo**	**Antecipação**	**Confiança**
Grupo 1	0.091	0.0	1.0	0.0	0.942	0.0171	1.0	1.0
Grupo 2	0.015	0.269	0.441	0.003	0.802	0.003	0.456	0.809
Grupo 3	0.143	0.0	1.0	1.0	0.952	0.079	0.860	0.988
Grupo 4	0.140	0.741	0.699	0.920	0.971	0.192	0.845	0.995

Tabela 6.4: Sentimentos associados a cada Grupo

Grupo	**Sentimentos**
Grupo 1	Submissão, Ansiedade, Esperança
Grupo 2	Cartas sem emoções associadas
Grupo 3	Desespero, Sentimentalismo e Submissão
Grupo 4	Amor e Sentimentalismo

Figura 6.3: Representação gráfica dos grupos representados pelas nuvens de palavras.

relata-se arrependimento por alguma ação ou postura da sua carreira artística além de submissão a Deus e a sua família como descrito no capítulo de referencial bibliográfico.

6.2 Análise das Pinturas e suas Correlações

No cenário de análise da complexidade de cada quadro e desenho observou-se que quanto mais para o final da carreira do artista mais complexa tornaram-se seus quadros e desenhos. Podemos correlacionar dois pontos interessantes:

- A saúde mental debilitada do artista, com diversas internações no hospital psiquiátrico identificado também através da análise de emoções como visualizado na Figura 6.2
- O aumento da média da complexidade de seus quadros no final da sua carreira artística que também foi a época de maior produção de Van Gogh.

A seguir visualizamos o gráfico das médias da complexidade de suas pinturas ao longo dos anos(Figura 6.4) , assim como a média de todos seus 1564 registros, incluindo desenhos e pinturas. (Figura 6.5). Analisando as médias das complexidades observamos, justamente o aumento da complexidade das obras ao longo da sua carreira da mesma maneira como o aumento do número de produções.

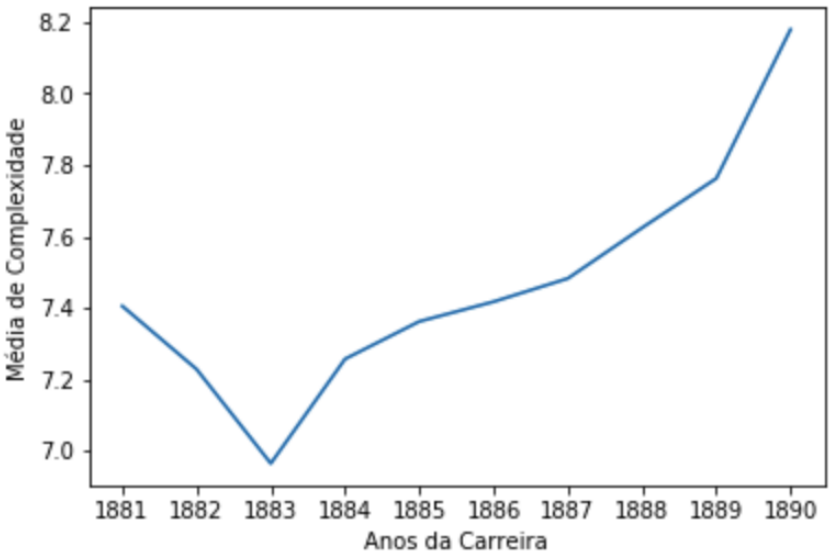


Figura 6.4: Gráficos da Complexidade somente das Pinturas

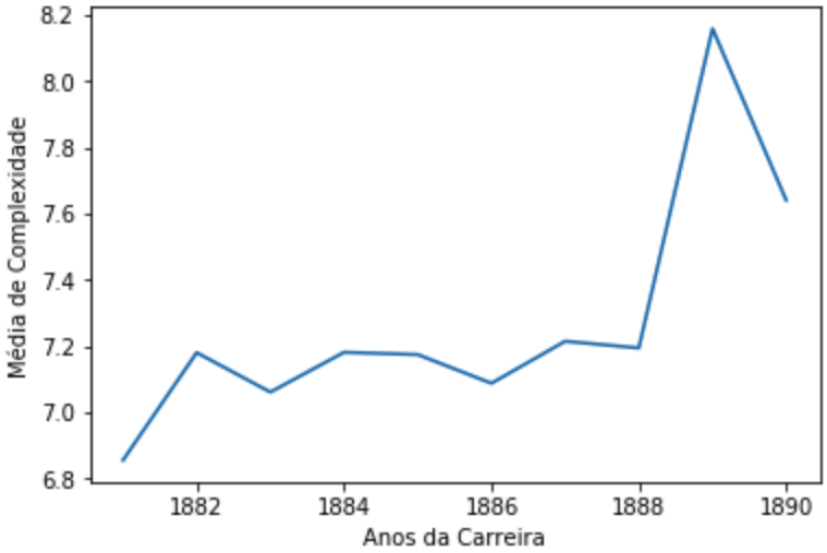


Figura 6.5: Gráficos da Complexidade Média das Pinturas e Desenhos

Além da complexidade, foi realizada uma análise das cores utilizadas pelo artista. Para cada ano foi realizado a quantização de cores de cada imagem, com o objetivo de extrair a paleta de cores para cada ano. Foi constatado nos primeiros anos o uso de cores mais escuras que ao longo de sua carreira, com influências de outros estilos, foi se tornando mais clara, ou seja, com o maior uso de cores variadas (Figura 6.6).

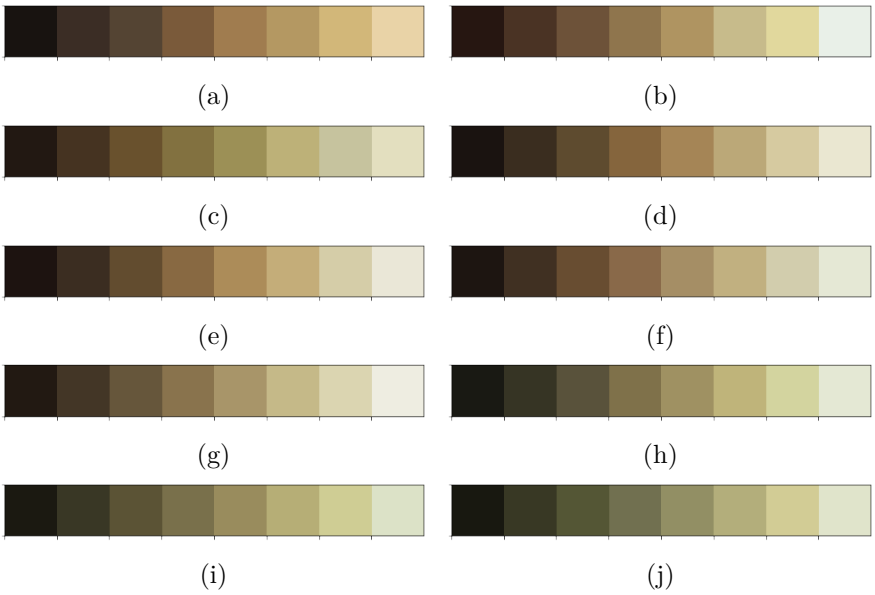
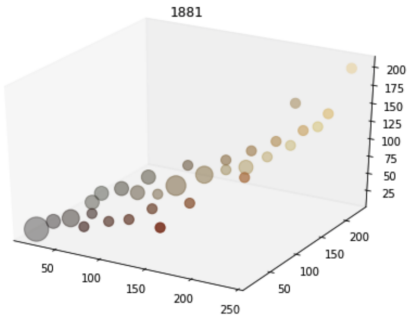


Figura 6.6: Paleta de cores média em referência aos anos 1881(a), 1882(b), 1883(c), 1884(d), 1885(e), 1886(f), 1887(g), 1888(h), 1889(i) e 1890(j).

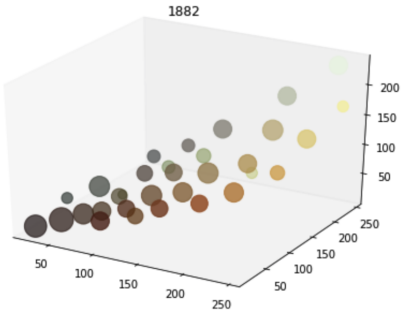
Essa correlação entre as cores utilizadas, e a complexidade de seus quadros são interessantes do ponto de vista da carreira do artista e um dos motivos pelos quais é reconhecido atualmente. . Prova-se uma relação entre os seus anos finais, internado no hospital psiquiátrico e a criação de suas obras mais complexas em diversos níveis, entre eles na variação de cores, proporcionando em obras famosas como "A noite estrelada"e seus famosos autorretratos.

Além da paleta de cores para cada ano, aferiu-se também a paleta de cores comparadas com a quantidade utilizada de cada cor, essa sendo apresentada na Figura 6.7. Percebemos também a correlação entre a concentração das cores utilizadas dentro do cubo RGB e o aumento da complexidade.

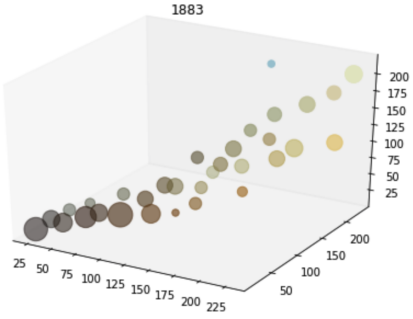
Ainda dentro da análise das cores, a esparsidade e concentração dos centrotos das cores foi calculada a partir da distância entre as cores utilizando ΔE definido no espaço de cor CIEL*a*b. Para cada ano, foi calculado a distribuição das distâncias entre os centroides utilizando a segmentação por 8, 16, 32 e 64 cores.



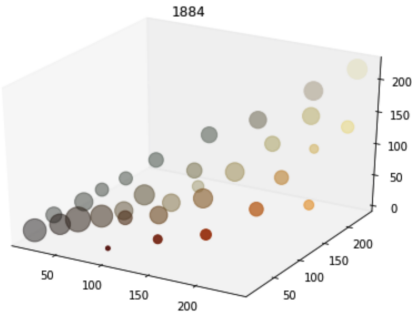
(a)



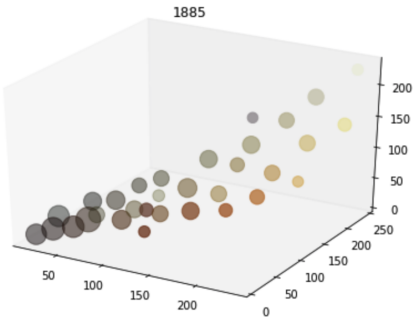
(b)



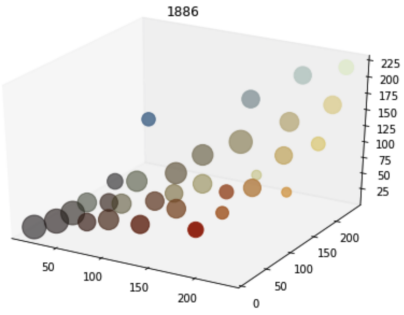
(c)



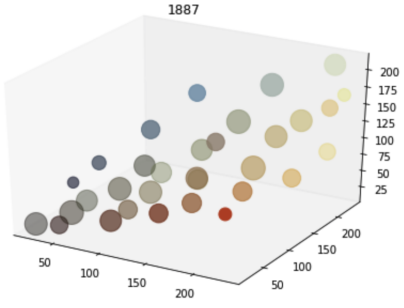
(d)



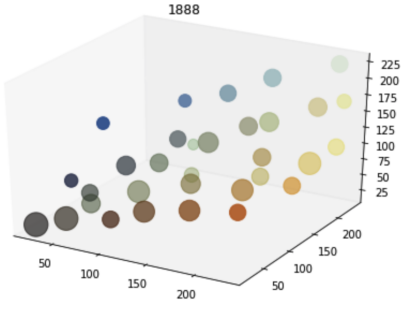
(e)



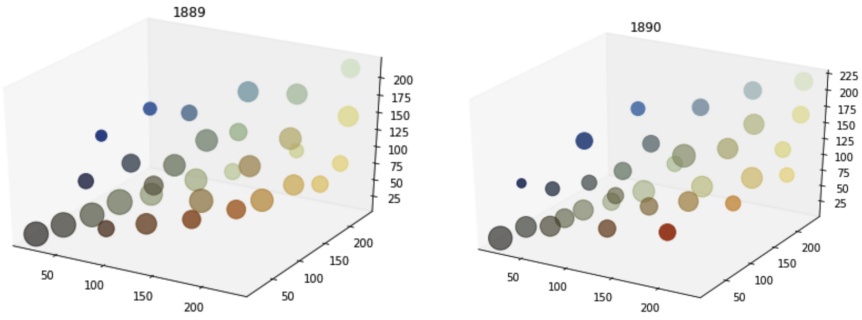
(f)



(g)



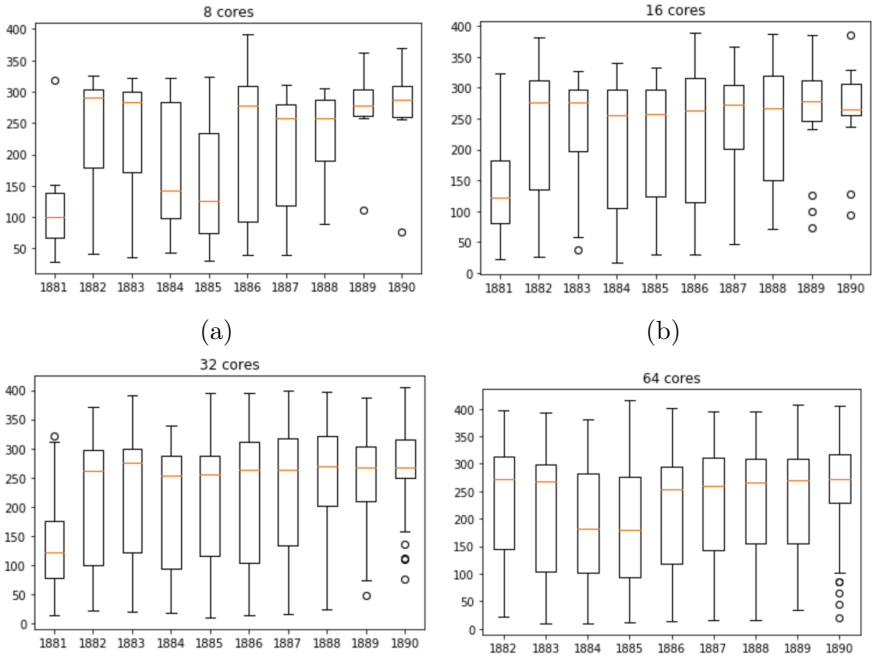
(h)



(i) (j)

Figura 6.7: Gráfico de cores em referência aos anos 1881(a), 1882(b), 1883(c), 1884(d), 1885(e), 1886(f), 1887(g), 1888(h), 1889(i) e 1890(j).

Dessa forma podemos comparar a variação das cores ao longo do tempo com a complexidade dos quadros de Van Gogh, achando relações interessantes também junto as cores(Figura 6.8).



(a) (b)

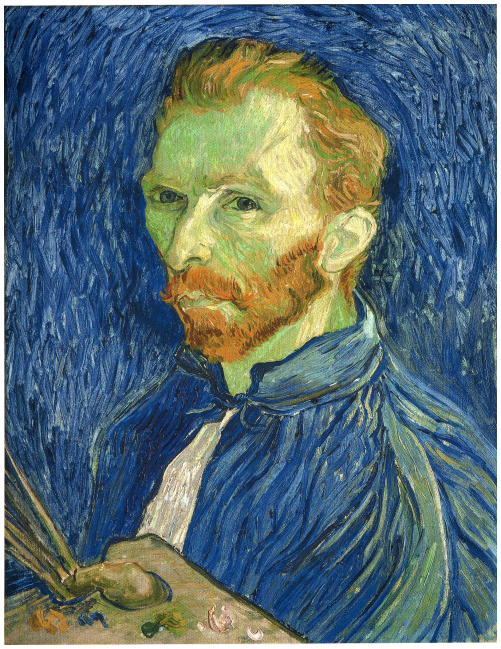
(c) (d)

Figura 6.8: Gráfico da distribuição das distância entre os centroídes para uma paleta de 8(a), 16(b), 32(c) e 64(d) cores.

(a) Cottage with Decrepit Barn and Stooping Woman, 1885

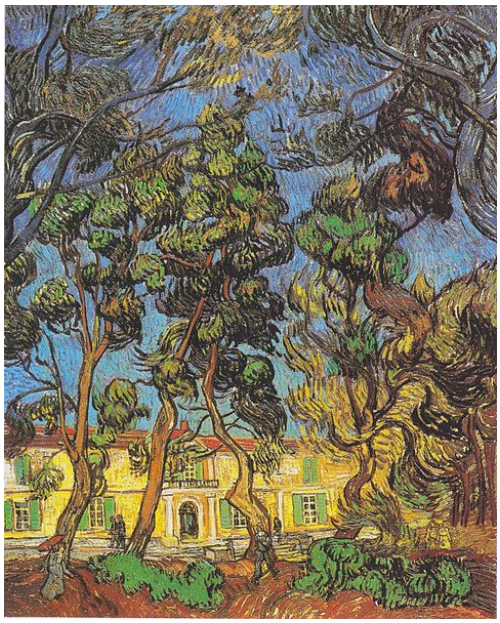
(b) The Pont du Carrousel and the Louvre, 1886

(c) Plaster Statuette of a Female Torso, 1887



(d) Self Portrait with Pallette, 1889

(e) Self Portrait with Bandaged Ear, 1889



(f) Trees in the garden of the Hospital Saint-Paul, 1888

Figura 6.9: Exemplo de quadros e desenhos de menor complexidade em (a), (b) e (c); e de maior complexidade em (d), (e) e (f)

Capítulo 7

Conclusão

Ferramentas que auxiliem na curadoria de arte, visam melhorar a análise da obra de um artista além de contribuir para o entendimento da história e das personalidades, como no caso de Vincent Van Gogh.

Sua vida conturbada e sua curta carreira artística impulsionaram ainda mais a concepção de suas pinturas com características atemporais e únicas. Além de entender sua vida, a identificação das emoções envolvidas nos ajudam a melhorar o entendimento sobre sua visão diferenciada do mundo e até mesmo enaltecer como sua saúde mental influenciou em sua obra.

Vemos também que com a avanço do poder de processamento e a disponibilidade de algoritmos de aprendizado de máquina podemos lidar com a interpretação de tarefas antes feitas exclusivamente pelo ser humano, como processamento de texto e de imagens como utilizados nos trabalhos [21, 27, 23, 28, 29, 24].

Esse trabalho teve como intuito analisar os dados textuais disponíveis nas cartas de Vincent Van Gogh de forma a identificar emoções e sentimentos intrínsecos em sua trajetória artística através de métodos de análise de emoções e algoritmos de classificação não supervisionados, como o *K-Means*. Assim como, verificar ao longo do tempo a complexidade de suas pinturas e desenhos com o auxílio de técnicas de processamento de imagens utilizando descritores de textura GLCM, informações sobre bordas, cores e os respectivos histogramas.

Como um trabalho exploratório, os resultados fazem parte de uma análise subjetiva que tem como principal propósito auxiliar no processo de curadoria da vida do artista de forma a tornar mais robusta e precisa, sendo validada a partir da correlação dos padrões encontrados, via a metodologia utilizada, e a biografia do artista. Os resultados encontrados como as emoções de submissão, desespero, ansiedade, esperança, sentimentalismo e amor possuem total relação com seu estado mental, a complexidade de suas pinturas e

sua variação de cores ao longo de sua carreira artística como apresentado no Capítulo 4, reforçam justamente o alcance dos objetivos estabelecidos anteriormente.

É definido como trabalho futuro a utilização de uma metodologia híbrida para aumentar a robustez dos resultados na parte de processamento de linguagem natural com o intuito melhorar a correlação entre as cartas e as emoções. Assim como uma análise de psicologia das cores no que diz respeito ao estudo de suas pinturas e desenhos. Outra proposta visa agrupar mais características junto a curadores de artes para aprimorar a metodologia adotada e melhorar a ferramenta desenvolvida.

Esse trabalho também contribui com uma base de dados pública das cartas, quadros e desenhos do artista Van Gogh, de forma organizada e disponível para futuros trabalhos e estudos na área de mineração de dados e apredizagem de máquina.

Referências

[1] Museum, Van Gogh: *Van gogh museum - meet vincent.* https://www.vangoghmuseum.nl/en/vincent-van-gogh-life-and-work, acesso em 2019-02-27. ix, 4, 5, 7, 8

[2] Plutchik, Robert: *The nature of emotions.* American Scientist, 89:344, janeiro 2001. ix, xi, 17, 18, 37

[3] Gonzalez, Rafael C. e Richard E. Woods: *Digital Image Processing (3rd Edition).* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2006, ISBN 013168728X. ix, 21, 22, 23, 24, 25

[4] Matlab: *Tutorial matalb - creating gray-level co-occurrence matrix.* https://www.mathworks.com/help/images/create-a-gray-level-co-occurrence-matrix.html, acesso em 2019-02-27. ix, 26

[5] Han, Jiawei, Micheline Kamber e Jian Pei: *Data mining concepts and techniques, third edition*, 2012, ISBN 0123814790. http://www.amazon.de/Data-Mining-Concepts-Techniques-Management/dp/0123814790/ref=tmm_hrd_title_0?ie=UTF8&qid=1366039033&sr=1-1. ix, 27, 28, 29

[6] De Botton, Alain e Armstrong, John: *Arte como terapia.* Editora Intrísica, 2014, ISBN 9788580575699. 3, 4

[7] Naifeh, Steven e White S., Gregory: *Van Gogh: A vida.* Companhia das Letras, 2014, ISBN 9788535921977. 4

[8] Kasabov, Nikola K.: *Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering.* MIT Press, Cambridge, MA, USA, 1st edição, 1996, ISBN 0262112124. 9, 10, 27

[9] Fogel, David: *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence*, volume 1. janeiro 1995, ISBN 978-0-7803-1038-4. 9

[10] Russell, S. e P. Norvig: *Artificial Intelligence: A Modern Approach.* Series in Artificial Intelligence. Prentice Hall, Upper Saddle River, NJ, third edição, 2010. http://aima.cs.berkeley.edu/. 9, 10, 13

[11] Mitchell, Ryan: *Web Scraping with Python: Collecting Data from the Modern Web.* O'Reilly Media, Inc., 1st edição, 2015, ISBN 1491910291, 9781491910290. 11

[12] Bird, Steven, Ewan Klein e Edward Loper: *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edição, 2009, ISBN 0596516495, 9780596516499. 9, 10, 12, 15

[13] Manning, Christopher D. e Hinrich Schütze: *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999, ISBN 978-0-262-13360-9. 9, 12, 15

[14] Jurafsky, Daniel e James H. Martin: *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edição, 2000, ISBN 0130950696. 13, 15, 16

[15] Ingersoll, Grant S., Thomas S. Morton e Andrew L. Farris: *Taming Text: How to Find, Organize, and Manipulate It*. Manning Publications Co., Greenwich, CT, USA, 2013, ISBN 193398838X, 9781933988382. 15, 16

[16] NLTK: *Nltk*. `https://www.nltk.org/`, acesso em 2019-02-27. 17

[17] Shelke, Nilesh, Shrinivas Deshpande e V. M. Thakare: *Approach for Emotion Extraction from Text*, páginas 661–669. março 2017. 19

[18] Manning, Christopher: *Natural language processing with deep learning -cs224n*. `http://web.stanford.edu/class/cs224n/slides/cs224n-2019-lecture01-wordvecs1.pdf`, acesso em 2019-02-27. 19

[19] Mikolov, Tomas, Kai Chen, Greg Corrado e Jeffrey Dean: *Efficient estimation of word representations in vector space*. CoRR, abs/1301.3781, 2013. `http://dblp.uni-trier.de/db/journals/corr/corr1301.htmlabs-1301-3781`. 19

[20] Duda, Richard O., Peter E. Hart e David G. Stork: *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2ª edição, November 2000, ISBN 0471056693. 27

[21] Ahuja, S. e G. Dubey: *Clustering and sentiment analysis on twitter data*. Em *2017 2nd International Conference on Telecommunication and Networks (TEL-NET)*, páginas 1–5, Aug 2017. 30, 64

[22] Folego, G., O. Gomes e A. Rocha: *From impressionism to expressionism: Automatically identifying van gogh's paintings*. Em *2016 IEEE International Conference on Image Processing (ICIP)*, páginas 141–145, Sep. 2016. 31

[23] Alshari, E. M., A. Azman, S. Doraisamy, N. Mustapha e M. Alkeshr: *Improvement of sentiment analysis based on clustering of word2vec features*. Em *2017 28th International Workshop on Database and Expert Systems Applications (DEXA)*, páginas 123–126, Aug 2017. 32, 64

[24] Viswanathan, Nitin: *Artist identification with concolutional neural networks*. Stanford University, 2017. 32, 64

[25] Madhoushi, Z., A. R. Hamdan e S. Zainudin: *Sentiment analysis techniques in recent works*. Em *2015 Science and Information Conference (SAI)*, páginas 288–291, July 2015. 33

[26] Chen, Y., J. Duan, Y. Zhu, X. Qian e B. Xiao: *Research on the image complexity based on neural network*. Em *2015 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 1, páginas 295–300, July 2015. 34

[27] Zhang, X. e Q. Yu: *Hotel reviews sentiment analysis based on word vector clustering*. Em *2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA)*, páginas 260–264, Sep. 2017. 64

[28] Çoban, Ö. e G. T. Özyer: *Word2vec and clustering based twitter sentiment analysis*. Em *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, páginas 1–5, Sep. 2018. 64

[29] Alshari, E. M., A. Azman, S. Doraisamy, N. Mustapha e M. Alkeshr: *Effective method for sentiment lexical dictionary enrichment based on word2vec for sentiment analysis*. Em *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, páginas 1–5, March 2018. 64