PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
ESCOLA POLITÉCNICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

MINERAÇÃO DE TEXTO APLICADA À IDENTIFICAÇÃO DE SENTIMENTOS E INTENÇÕES

IGOR FERREIRA DE JESUS PEREIRA

GOIÂNIA
2022

IGOR FERREIRA DE JESUS PEREIRA

MINERAÇÃO DE TEXTO APLICADA À IDENTIFICAÇÃO DE SENTIMENTOS E INTENÇÕES

Trabalho de Conclusão de Curso apresentado à Escola de Ciências Exatas e da Computação, da Pontifícia Universidade Católica de Goiás, como parte dos requisitos para a obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Sibelius Lellis Vieira
Banca examinadora: Prof. Me. André Luiz Alves
Prof. Me. Gustavo Siqueira Vinhal

GOIÂNIA
2022

IGOR FERREIRA DE JESUS PEREIRA

MINERAÇÃO DE TEXTO APLICADA À IDENTIFICAÇÃO DE SENTIMENTOS E INTENÇÕES

Trabalho de Conclusão de Curso aprovado em sua forma parcial pela Escola de Ciências Exatas e da Computação, da Pontifícia Universidade Católica de Goiás, para obtenção do título de Bacharel em Ciência da Computação, em _____/_____/_____.

_____

Orientador: Prof. Dr. Sibelius Lellis Vieira

_____

Prof. Me. André Luiz Alves

_____

Prof. Me. Gustavo Siqueira Vinhal

GOIÂNIA

2022

AGRADECIMENTO

Agradeço primeiramente a Deus, por me dar forças para perseverar e vencer todos os desafios que tive em minha jornada acadêmica.

Aos meus pais, Maria Celma Ferreira e Uelton de Jesus Pereira, pelo amor incondicional, apoio e sacrifícios para me tornarem quem eu sou.

A minha esposa Adrya Viera de Oliveira, a qual sempre me apoia e me auxilia em tudo. Estando sempre presente ao meu lado, nos momentos bons e ruins.

Ao professor Dr. Sibelius Lellis Vieira, por todo apoio, compreensão, orientação e correções nesta jornada. Sem ele, este trabalho não seria possível.

E, por fim, a todos que contribuíram de alguma forma para a conclusão desta jornada.

RESUMO

O ser humano está rodeado de informação e a tecnologia revolucionou a velocidade e facilidade de acesso a ela. O que antes estava apenas em livros, jornais, revista e outros meios, se encontra hoje disponível, na palma da mão. Em muitos casos, essas informações podem, por exemplo, serem extraídas de músicas. A música pode ser apresentada de forma escrita, oral ou audiovisual e tem relação longa com a humanidade, sendo uma das formas de manifestações artísticas e culturais mais antigas. Pensando nas letras de músicas, esse conjunto de textos é muito amplo e rico em informações, havendo vários estilos e gêneros, podendo contar diferentes histórias, expressar diversos sentimentos etc. O benefício da Mineração de Texto (*Text Mining*) se dá pela grande quantidade de informações valiosas contidas nos textos. A proposta geral deste trabalho é aplicar técnicas de mineração de textos em um conjunto de letras musicais a fim de obter padrões de interesse através de um estudo de caso sobre informações contidas em letras de músicas e a intenção emocional existente nas palavras que compõem seu conteúdo. O projeto é classificado de acordo com seus objetivos de natureza exploratório e experimental com abordagem predominantemente qualitativa, recorrendo a análises quantitativas utilizado do método de estudo de caso. Como resultado da aplicação dos conceitos de descoberta de conhecimento e as técnicas de mineração de texto, foi possível identificar o conteúdo e o sentimento envolvido no conjunto de dados utilizados de modo a observar sentimentos negativos e egocêntricos, tais como: não quero, não vou, não sei, não vai, pra mim entre outros.

Palavras-chave: *Text Mining*, conjunto de dados, letras de músicas, frequência de termos, análise de sentimentos, unigrama, bigrama.

ABSTRACT

The human being is surrounded by data and information technology has revolutionized its access and efficient processing. What was once presente in printed materials like books, newspapers, magazines and so on, is now available, at the front of the hand. Sometimes, this information can, for example, be extracted from music, lyrics and songs. Music can be presented in written, spoken or audio-visual form and has a long relationship with the human society, being one of the oldest artistic and cultural manifestations. The benefit of Text Mining is given by the large amount of information contained in the texts. The application of this work is to apply to a set of musical lyrics a series of text mining techniques, in order to obtain patterns of interest regarding word frequency, sentimento anaysis, relationship between words and the most prevalente analysis. This work is classified as a case study, descriptive, exploratory and experimental one. As a result of applying knowledge discovery concepts and text mining techniques, it was possible to identify the content and sentiment involved in the dataset used in order to observe negative and egocentric feelings, such as: I don't want, I won't, I don't know, won't, for me, among others.

Keywords: Text Minig, dataset, song lyrics, terms frequency, sentiment analysis, unigrams, bigrams.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS

AFFIN - Finn Ärup Nielsen

AMD - *Advanced Micro Devices Inc*

API - *Application Programming Interface*

AS - Análise de Sentimento

BING - Bing Liu

CD - *Compact Disc*

CSV - *Comma-separated values*

DCDB - Descoberta de conhecimento em bases de dados

GHz - Gigahertz

IDE - *Integrated Development Environment*

KDD - *Knowledge Discovery in Databases*

MP3 - MPEG-1/2 *Audio Layer* 3

OC - Organização do Conhecimento

PLN - Processamento da Linguagem Natural

RAM - *Random Access Memory*

TM - *Text Mining*

XML - e*Xtensible Markup Language*

Sumário

1. INTRODUÇÃO

1.1 Contextualização

A informação tem um papel fundamental na forma como a sociedade se comunica, age e pensa. Como conjunto de textos, a música é muito ampla e rica em informações, havendo vários estilos e gêneros, podendo contar diferentes histórias, expressar diversos sentimentos etc.

A música é uma forma de arte que possui uma relação longa com a humanidade, sendo uma das formas de manifestações culturais mais antigas. Ela é reconhecida por muitos pesquisadores como uma modalidade que desenvolve a mente humana, promove o equilíbrio, proporcionando um estado agradável de bem-estar, facilitando a concentração e o desenvolvimento do raciocínio, em especial em questões reflexivas voltadas para o pensamento (GODOY, 2019).

Técnicas computacionais podem ser usadas para identificar tendências e padrões musicais através da extração de padrões de dados, podendo contribuir para o ecossistema da indústria da música. Essas técnicas são bastante utilizadas em diversas áreas do conhecimento, como: Inteligência Artificial, *Data Science* (Ciência de dados), *Machine Learning* (Aprendizado de máquina) e análise estatísticas. Elas consistem na extração das características de áudio (conteúdo) ou letra (contexto), sendo preferível as letras por exigirem menor custo computacional e apresentar melhores resultados. Por meio da extração de conhecimento e análise baseadas em letras é possível identificar a intenção emocional presentes em músicas e explicar como os subgêneros diferem (JUNIOR; ROSSI; LOBATO, 2019).

1.2 Justificativa

Segundo Tan (1999) *Text Mining* é o processo de obtenção de informações importantes de bases textuais não estruturadas, mas pode também ser visto como uma extensão da Mineração de Dados (*Data Mining*), que é a extração de conhecimento de bases de dados estruturadas.

Portanto, nesse contexto, o benefício da Mineração de Texto (*Text Mining*) se dá pela grande quantidade de informações valiosas contidas nos textos, como tendências, anomalias e padrões de comportamento que podem ser utilizados (BERRY; KOGAN, 2010).

1.3 Objetivo

1.3.1 Objetivo geral

A proposta geral deste trabalho é aplicar técnicas de mineração de textos em um conjunto de letras musicais a fim de obter padrões de interesse através de um estudo de caso sobre informações contidas em letras de músicas e a intenção emocional existente nas palavras que compõem seu conteúdo.

1.3.2 Objetivos específicos

Para se atingir o objetivo geral, propõem-se os seguintes objetivos específicos:

- Realizar a formatação dos dados;
- Realizar análise de frequência dos termos;
- Gerar a nuvem de palavras;
- Realizar a análise de sentimentos.

1.4 Estrutura do trabalho

Esse trabalho apresenta-se estruturado em cinco capítulos, sendo o primeiro capítulo esta introdução. O segundo capítulo aborda o referencial teórico que contém todo o fundamento relacionado ao tema desenvolvido, como a música e sua aplicação a Ciência da Dados para descoberta de conhecimento e as técnicas de mineração de texto aplicadas a letras de música. O terceiro capítulo descreve os materiais e métodos utilizados para realização dos experimentos. O quarto capítulo apresenta os resultados obtidos das análises dos termos unigramas e bigramas contidas no conjunto de dados. O quinto capítulo traz as considerações finais e descreve possíveis trabalhos futuros.

2. REFERENCIAL TEÓRICO

Este capítulo tem como objetivo apresentar o embasamento teórico e características da mineração de texto em letras de músicas.

2.1 Textos da internet

Segundo Freitas (2004) o processo de coleta de dados é limitado pelo custo, tempo e intensidade de trabalho. Porém essas limitações podem ser quebradas com o uso da internet, uma vez que ela proporciona um ambiente de grande facilidade tanto para produção, aquisição e difusão de informação, quanto do tratamento dos dados necessários. A informação tem um papel fundamental na forma como a sociedade se comunica, age e pensa, sendo considerada uma das tecnologias de maior influência na vida do ser humano.

Existem três tipos de dados obtidos pela internet: os estruturados, semiestruturados e não estruturados. Os dados estruturados têm como características esquemas rígidos e adequados para o formato de tabelas. Os dados semiestruturados são aqueles que possuem uma estrutura pré-definida, porém não com o mesmo rigor dos dados estruturados. Normalmente apresentam apenas um meio de marcação dos dados, como é o caso dos arquivos no formato XML (*eXtensible Markup Language*). Na classe de dados não estruturados estão inclusos os vídeos, imagens, e alguns formatos de textos (MARQUESONE, 2017).

Segundo Berry e Kogan (2010) com o avanço da tecnologia da informação, notou-se que em grande parte das redes sociais armazenam dados semiestruturado e não estruturados. Esses dados possuem informações valiosas, como por exemplo: tendências, anomalias e padrões de comportamento que podem ser utilizados.

A música vista como objeto informacional é um tema complexo quando aplicado ao campo da organização do conhecimento (OC), principalmente quando se almeja modelar um domínio de forma automatizada. Uma das técnicas que vem despontando como promissora neste setor é a Análise de Sentimento (AS) ou Mineração de Opinião, técnica derivada da inteligência artificial para identificar opiniões e emoções em textos, avaliando-as como positivas ou negativas por meio do processamento automático da linguagem natural (SOUZA; CAFÉ, 2018).

2.1.1 A música

A música é uma forma de arte que possui uma relação longa com a humanidade, sendo uma das formas de manifestações culturais mais antigas. Ela é constituída de um trabalho com a harmonia entre os sons, ritmos, melodias, letra e a voz (AIDAR, 2019).

Segundo Esteves (2019) esses sons podem ser produzidos pelo corpo qualquer, transmitidas por um meio (sólido, líquido ou gasoso) por meio de propagação de frequências regulares ou não, captadas pelos ouvidos e interpretadas pelo cérebro. Os sons podem ser produzidos, por exemplo: pelo corpo humano (voz e percussão corporal), por instrumentos musicais (acústicos, como um piano, sax ou violão), elétricos (como uma guitarra, baixo, teclado em um amplificador) ou digitais (como um sintetizador). Podem também ser, ainda, uma gravação (como um disco de vinil, um CD (Compact Disc), ou arquivo MP3 (MPEG-1/2 *Audio Layer* 3)), porém assim como qualquer outro som, de qualquer natureza, são organizados de uma maneira lógica.

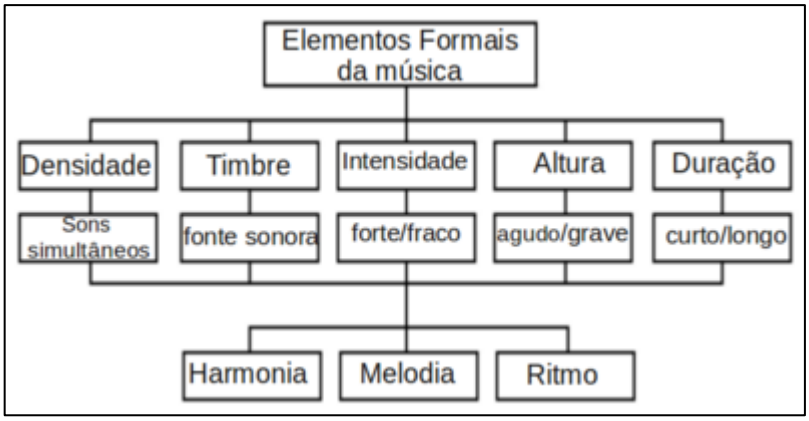
O som possui cinco propriedades:

- Intensidade: é a força do som, também chamada de sonoridade, sendo ela responsável por permitir ao ouvinte distinguir se o som é fraco (baixa intensidade), ou se o som é forte (alta intensidade) e ela está relacionada à energia de vibração da fonte que emite as ondas sonoras;
- Duração: é o tempo que o som permanece em um meio, podendo ser curto ou longo;
- Altura: é meio de distinguir um som agudo (fino, alto), de um grave (grosso, baixo). A altura de um som musical depende do número de vibrações, sendo as vibrações rápidas responsáveis pela produção de sons agudos e as lentas pelas pela produção de sons graves. Essas vibrações é que definem cada uma das notas musicais: dó, ré, mi, fá, sol, lá, si; a velocidade da onda sonora determina a altura do som, por isso, cada nota sua frequência (número de vibrações por segundo);
- Timbre: é a propriedade que distingue a qualidade do tome ou voz de um instrumento ou cantor. Em que cada objeto ou material possuí um timbre que é único, assim como cada pessoa possui um timbre próprio de voz, tão individual quanto as impressões digitais;

- Densidade: é qualidade que estabelece um maior ou menor número de sons simultâneos.

A Figura 1 ilustra os elementos formais da música.

Figura 1- Elementos Formais da música



Fonte: Adaptada Secretaria da Educação do Paraná (2012)

A música além de suas propriedades sonoras, possui também os elementos de estruturação que permite com que o som seja agradável de se ouvir, sendo eles:

- Melodia: é uma sequência de sons em intervalos irregulares que caminha entre o ritmo, sendo a parte mais desatava da música e a parte que fica a cargo do Cantor, ou de um instrumento de destaque como o Sax ou de um solo de guitarra;
- Harmonia: é a combinação dos sons ouvidos simultaneamente, sendo o agrupamento agradável de sons;
- Ritmo: é o que age em função da duração do som, definindo quanto tempo cada parte da melodia continuará à tona.

Para que a música faça sentido é preciso que esses elementos sejam organizar a escuta, normalmente são organizados em “seções”, ou “partes". Esse tipo de organização é conhecimento como Forma, dividindo a música em “introdução”, “verso” e “refrão”.

Com a junção de todos esses elementos é possível categorizar as músicas que compartilham elementos em comum, determinando o que se chama de gêneros musicais. Dentre destes, os gêneros são definidos também a partir dos elementos contidos na música, como:

- Instrumentação;
- Letra;
- Função;
- Estrutura;
- Contextualização.

A Figura 2 ilustra alguns dos gêneros musicais que podem conter em seu conteúdo diferentes histórias e expressar diversos sentimentos.

Figura 2 - Gêneros musicais

Fonte: Adaptada de Gêneros musicais e suas definições (2014)

2.1.2 Ciência de dados aplicada a música

A música pode-se materializar de diversas formas: partituras, registros sonoros, letras de músicas entre outras estruturas. Há autores que acreditam serem as letras uma importante fonte de representação sobre música em sua dimensão emocional (SOUZA; CAFÉ, 2018).

Para realizar a análise de sentimentos de um texto pode-se utilizar o significado semântico das palavras contidas nele. Dessa forma ao utilizar a letra, pensando na ideia que apenas lendo a letra de uma música é possível identificar o gênero, humor e temática dela.

Segundo os Souza e Café (2018) os três maiores obstáculos na análise de sentimento por meio das letras de músicas:

1. Apenas a análise de uma série de trechos de uma letra, não é fator determinante para todo o contexto;
2. Letras são subjetivas e podem atribuir a um léxico, um sentimento positivo ou negativo mesmo sendo não determinado;
3. Letras podem expressar emoções que são contrarias aos fatos.

Essa dificuldade ocorre devido a música conter dois atores principais, sendo o compositor e o ouvinte que tem seus próprios pensamentos e ideias, defendendo seus pontos de vista.

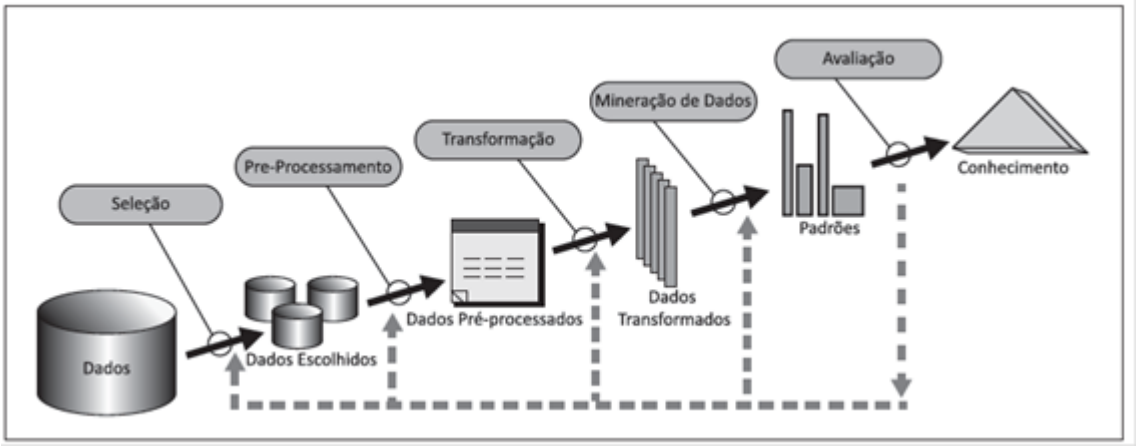
Observar os aspectos morfológicos, sintáticos e semânticos das letras musicais permitem identificar em seu conteúdo linguísticos e melódicos, pensamentos e ideias que são transmitidos por meio de mensagens. Essas mensagens podem ser de caráter efetivo, influenciar e serem influenciadas por valores sociais, culturais e políticos.

2.2 Descoberta de conhecimento em bases de dados (DCDB)

Devido o crescente número de informações que são geradas e armazenadas na atualidade, a análise dessas informações torna-se inviável de forma manual, sendo necessário o uso de ferramentas para automatizar e auxiliar neste processo.

Segundo Fayyad (1996), a Descorberta de conhecimento em base de ados, em português, ou chamado KDD (*Knowledge Discovery in Databases*) se refere ao processo não trivial de identificar a partir de dados armazenados em banco de dados novos padrões que sejam válidos, antes não conhecidos, e que são potencialmente úteis e compreensivos. Isso possibilita ter um melhor entendimento de um problema ou procedimento de tomada de decisão. A Figura 3 ilustra o processo de extração de informação em meio a um conjunto de dados.

Figura 3 - Processo *Knowledge Discovery in Databases* (KDD)



Fonte: Adaptada de FAYYAD et. al. (1996)

O processo de KDD é formatado por cinco etapas, sendo elas: seleção, pré-processamento, transformação, mineração de dados e interpretação dos resultados. Estas etapas são distribuídas em 3 (três) principais grupos: pré-processamento (seleção de dados, limpeza dos dados e tratamento de dados); processamento (mineração de dados); pós-processamento (interpretação).

Segundo Morais e Ambrósio (2007) na etapa de seleção tem como objetivo localizar e filtrar os documentos com informações relevantes para o assunto que está sendo abordado, visando eliminar inconsistências, registros incompletos e valores errados.

Na etapa de Pré-processamento é um processo de estruturar os dados, analisando se existem redundâncias, dependências entre as variáveis e valores conflitantes.

Na transformação dos dados alguns algoritmos trabalham apenas com valores numéricos e outros apenas com valores categóricos. Com isso, é necessário transformar os dados para o mesmo o valor (CAMILO; SILVA, 2009).

Depois o que os textos selecionados foram transformados para os dados estruturados, segue-se para o estágio de mineração de dados, onde tenta-se descobrir um padrão nos dados selecionados (MORAIS; AMBRÓSIO, 2007).

No processo de interpretação, mostra-se os padrões encontrados na fase de mineração. Essa interpretação é feita usualmente com linguagem natural (MORAIS; AMBRÓSIO, 2007).

Para Freitas (2018) o KDD refere-se a todo o processo de descoberta do conhecimento útil de dados, formatando para um padrão válido, enquanto a mineração de dados (*Data Mining*) é uma etapa responsável pela seleção dos métodos a serem utilizados, buscando por padrões de interesse. A mineração de texto (*Text Mining*) é vista como uma extensão do *Data Mining*, mas com objetivo na extração de informação útil em documentos de textos não estruturados.

2.3 Mineração de textos (*Text Mining*)

Segundo Kwartler (2017) "a mineração de texto é o processo de destilar percepções acionáveis do texto, com o objetivo final de auxiliar nas tomadas de decisão", apresentando assim, duas abordagens para o processo de mineração em bases textuais, a Análise Estatística e Análise Semântica.

A Análise Estatística trabalha com a frequência de aparição de cada termo em uma frase, não se importando com o contexto inserido. Pode ser representada por um gráfico de barras ou nuvem de palavras. A nuvem de palavras disponibiliza uma visualização dos termos mais frequentes dentro do conjunto analisado, de forma que quanto maior tamanho da fonte, maior a sua frequência.

A Análise Semântica por sua vez, tem seu foco na funcionalidade dos termos, através do significado morfológico, sintático, semântico, pragmático, conforme o Processamento da Linguagem Natural (PLN).

O PLN é um conjunto de técnicas computacionais para analisar e representar ocorrências naturais de texto em um ou mais níveis de análise linguística, com o objetivo de se alcançar um processamento de linguagem similar ao humano para uma série de tarefas ou aplicações. O PLN lida com diversos elementos linguísticos e estrutura gramatical, sendo um processo complexo, paralelo à complexidade da linguagem natural (LIDDY, 2001).

Segundo Chowdhury (2003) ambas são utilizadas na análise de sentimentos em bases textuais, porém podem ser utilizadas de forma independente ou em conjunto.

2.3.1 Técnicas de mineração de textos

Alguns exemplos de técnicas utilizadas na mineração de textos são: tokenização; formatação dos dados; análise de frequência das palavras; nuvem de palavras; análise de sentimentos; correção entre palavras: n-gramas (SILGE; ROBINSON, 2017).

2.3.2 Formatação dos dados

Segundo Single e Robinson (2017) uma maneira poderosa de tornar o manuseio de dados mais fácil e eficaz são os princípios de dados organizados, e isso mesmo trabalhando com textos. Conforme descrito por Wickham (2014) os dados organizados têm uma estrutura específica em que: cada variável é uma coluna; cada observação é uma linha; cada tipo de unidade observacional é uma tabela.

Dessa forma, o formato do texto é definido como organizado sendo ele uma tabela com um token por linha. O token é uma unidade significativa do texto, como uma palavra que estamos interessados em usar na análise, e a tokenização é o processo de dividir o texto em tokens. Essa estrutura de token por linha contrasta com as maneiras com as quais o texto geralmente é armazenado nas análises, sendo como *strings* ou em matriz de termos do documento que aplicação de aprendizagem de máquina. Para a mineração de texto organizada, o *token* armazenado em cada linha geralmente é uma única palavra, mas também pode ser um n-grama, uma frase ou um parágrafo.

Os textos frequentemente utilizados em abordagens de mineração de texto são armazenados de seguintes formas:

- *Strings*: vetores de caracteres, dentro do R, e muitas vezes os dados de textos são lidos primeiro na memória neste formato;
- *Corpus*: esses tipos de objetos geralmente contêm *strings* brutas contendo metadados e detalhes adicionais;
- Matriz de termos do documento (*Document-term matrix*): esta é uma matriz espalhada que descreve uma coleção (ou seja, um corpus) de documentos com uma linha para cada documento e uma coluna para cada termo. O valor na matriz geralmente é a contagem de palavras.

Não é esperado que o usuário mantenha os dados de texto em um formato organizado o tempo todo durante a análise, por isso são utilizadas técnicas e ferramentas, populares na mineração de texto para limpar e manter os dados organizados.

2.3.3 Análise de frequência

Segundo Freitas (2018) uma tarefa comum e importante na mineração de texto é observar as frequências de palavras, pois através dela é possível descobrir assuntos que estão presentes na base de dados, além de avaliar se a análise está seguindo o caminho coerente visando as palavras de interesse.

Para este processo é importante que os dados já tenham sido formatados e estruturados. Removendo os números, os pontos, os espaços em branco, as palavras chamadas: *stopwords* que são palavras que não agregam um sentido específico na frase e são utilizadas apenas para ligar uma palavra a outra. A Figura 4 ilustra a frequência de palavras que mais estão presentes dentro de uma base de dados.

Figura 4 - Frequência de palavras (*Word Frequencies*)

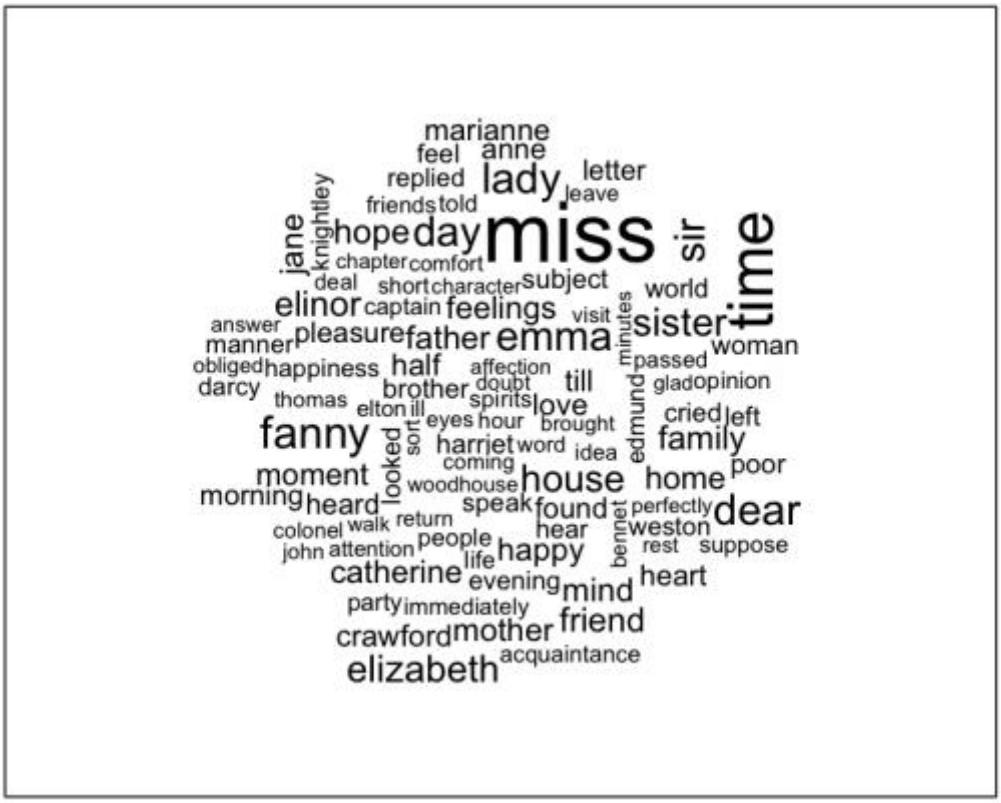
```
## # A tibble: 11,769 × 2
##      word     n
##     <chr> <int>
## 1    time   454
## 2  people   302
## 3    door   260
## 4   heard   249
## 5   black   232
## 6   stood   229
## 7   white   222
## 8    hand   218
## 9    kemp   213
## 10   eyes   210
## # ... with 11,759 more rows
```

Fonte: Adaptada de SILGE e ROBINSON (2017)

2.3.4 Nuvem de palavras

A da nuvem de palavras se trata de uma apresentação gráfica que lista de forma hierárquica as palavras de maior frequência dentro da base de dados. A Figura 5 ilustra uma nuvem de palavras.

Figura 5 – Nuvem de palavras (*Wordclouds*)



Fonte: Adaptada de SILGE e ROBINSON (2017)

2.3.5 Análise de sentimentos

Segundo Single e Robinson (2017) quando leitores abordam um texto, usa-se a compreensão da intenção emocional das palavras para inferir se uma seção do texto é positiva, negativa, ou se apresenta uma característica de outra emoção mais sutil, como: raiva, antecipação, nojo, medo, alegria, tristeza, surpresa ou confiança. Uma maneira de analisar o sentimento de um texto é considerar o texto com uma combinação de suas palavras individuais e o conteúdo do sentimento de todo o texto como a soma do conteúdo do sentimento de cada palavra.

Existem vários métodos de dicionários para avaliar a opinião ou emoção do texto. Três são comumente utilizados na mineração de texto:

- AFINN: atribui palavras como uma pontuação de que varia de -5 e 5, sendo os com pontuações negativas indicando sentimento negativo e pontuações positivas indicando sentimento positivo; A Figura 6 ilustra um conjunto de palavras e sua respectiva pontuação quanto ao sentimento atribuído a ela.

Figura 6 – Análise de sentimento AFINN

```
get_sentiments("afinn")

## # A tibble: 2,476 × 2
##          word score
##         <chr> <int>
## 1     abandon    -2
## 2   abandoned    -2
## 3    abandons    -2
## 4    abducted    -2
## 5   abduction    -2
## 6  abductions    -2
## 7       abhor    -3
## 8    abhorred    -3
## 9   abhorrent    -3
## 10     abhors    -3
## # ... with 2,466 more rows
```

Fonte: Adaptada de SILGE e ROBINSON (2017)

- BING: categoriza as palavras de forma binária em positivas e negativas; A Figura 7 ilustra um conjunto de palavras e suas respectivas categorias.

Figura 7 – Análise de sentimento BING

```
get_sentiments("bing")

## # A tibble: 6,788 × 2
##           word sentiment
##          <chr>     <chr>
## 1      2-faced  negative
## 2      2-faces  negative
## 3           a+  positive
## 4     abnormal  negative
## 5      abolish  negative
## 6   abominable  negative
## 7   abominably  negative
## 8    abominate  negative
## 9  abomination  negative
## 10       abort  negative
## # ... with 6,778 more rows
```

Fonte: Adaptada de SILGE e ROBINSON (2017)

- NRC: categoriza as palavras de forma binária (“sim” /” não”) para sentimentos positivos, negativos, raiva, antecipação, desgosto, medo, alegria, tristeza, surpresa e confiança. A Figura 8 ilustra um conjunto de palavras e o respectivo sentimento atribuído a ela.

Figura 8 – Análise de sentimento NRC

```
get_sentiments("nrc")

## # A tibble: 13,901 × 2
##           word sentiment
##          <chr>     <chr>
## 1       abacus     trust
## 2      abandon      fear
## 3      abandon  negative
## 4      abandon   sadness
## 5    abandoned     anger
## 6    abandoned      fear
## 7    abandoned  negative
## 8    abandoned   sadness
## 9  abandonment     anger
## 10 abandonment      fear
## # ... with 13,891 more rows
```

Fonte: Adaptada de SILGE e ROBINSON (2017)

2.3.6 Correlação entre palavras

Segundo Single e Robinson (2017) essa análise se baseia nas relações entre palavras, seja examinando quais palavras tendem a seguir outras imediatamente, ou que tendem a coocorrer dentro do mesmo documento. Atribuindo ao *token* como sendo: um par ou uma sequência de dois elementos adjacentes, chamado de "**bigramas**"; conforme ilustração apresentada na Figura 9; um trio ou uma sequência de três elementos adjacentes, chamado de "**trigramas**"; conforme ilustração apresentada na Figura 10; e uma relação chamado de "**n-gramas**", onde **n** é número de elementos adjacentes, podendo este conter vários elementos.

Figura 9 – Correlação entre duas palavras (bigramas)

```
## Source: local data frame [33,421 x 3]
## Groups: word1 [6,711]
##
##      word1     word2     n
##      <chr>     <chr> <int>
## 1      sir    thomas   287
## 2     miss  crawford   215
## 3  captain wentworth   170
## 4     miss woodhouse   162
## 5    frank churchill   132
## 6     lady   russell   118
## 7     lady   bertram   114
## 8      sir    walter   113
## 9     miss   fairfax   109
## 10 colonel   brandon   108
## # ... with 33,411 more rows
```

Fonte: Adaptada de SILGE e ROBINSON (2017)

Figura 10 – Correlação entre três palavras (trigramas)

```
## Source: local data frame [8,757 x 4]
## Groups: word1, word2 [7,462]
##
##        word1     word2     word3     n
##        <chr>     <chr>     <chr> <int>
## 1       dear      miss woodhouse    23
## 2       miss        de    bourgh    18
## 3       lady catherine        de    14
## 4  catherine        de    bourgh    13
## 5       poor      miss    taylor    11
## 6        sir    walter    elliot    11
## 7        ten  thousand    pounds    11
## 8       dear       sir    thomas    10
## 9     twenty  thousand    pounds     8
## 10   replied      miss  crawford     7
## # ... with 8,747 more rows
```

Fonte: Adaptada de SILGE e ROBINSON (2017)

2.3.7 Software R Studio

O R Studio é um *software* código aberto (*open-source*) que contém vários recursos para análise de *Text Minig*, sendo uma IDE (*Integrated Development Environment*), ou ambiente de desenvolvimento integrado para a linguagem R, uma linguagem de programação dinamicamente tipada, orientada a objetos que possui um amplo conjunto de pacotes e funções acessíveis para aplicações em diversas áreas do conhecimento, como: Ciência de dados (*Data Science),* Aprendizado de máquina (*Machine Learning*) e análise estatísticas.

2.4 Estudos correlatos

Há vários trabalhos relevantes disponíveis relacionados a mineração de texto, sendo o trabalho de Freitas (2018) um destes, no qual realizou um estudo com objetivo de coletar textos da rede social *Twitter* voltados as eleições para presidente do Brasil em 2018. Com isso, aplicar técnicas de mineração de texto, a fim traçar um perfil de interesse e opinião dos usuários referentes ao tema proposto.

Café e Souza (2017), estabelece em seu trabalho a música como objeto de estudo para aplicação de Análise de Sentimento, uma técnica da mineração de texto, a fim de identificar opiniões e emoções em textos contidos nas letras das músicas, avaliando-os como positivas ou negativas. Fornecendo em seu desenvolvimento, embasamento teórico para relevância do uso da música como objeto de estudo, coleta e construção do conjunto de dados, além da aplicação da análise de sentimento sobre o conjunto de dados, estabelecendo o sentimento baseado nas palavras contidas no texto e não na expressão artística.

3. MATERIAIS E MÉTODOS

Neste capítulo são apresentadas as ferramentas e metodologias utilizadas no desenvolvimento do trabalho.

3.1 Materiais

Durante a realização dos experimentos propostos foram selecionados ferramentas e materiais com propósitos específicos para determinada função. O conjunto de dados com as letras das músicas no formato CSV (*Comma-separated values*) para análise, o Excel para verificação e correção ortográfica do conjunto de dados. A análise dos dados é realizada com o *Software* R Studio.

3.1.1 Equipamentos e hardware

O equipamento de *hardware* utilizado durante a relação dos experimentos foi um desktop com processador AMD Ryzen 5 3600 3.6 Giga-hertz (GHz), oito Giga bytes de memória *Random Acess Memory* (RAM), Placa de vídeo RX 580 de oito Giga bytes de memória *Random Acess Memory* (RAM) e sistema operacional Windows 10, 64 bits, com os *softwares* necessários instalados e configurados.

3.1.2 Conjunto de dados

O conjunto de dados utilizado para análise deste trabalho foi disponibilizado de forma gratuita pelo site *Kaggle*. Ele se encontra disponível no endereço: https://www.kaggle.com/neisse/scrapped-lyrics-from-6-genres?select=lyrics-data.csv, contendo as letras de músicas extraídas do site brasileiro Vagalume pelo Anderson Neisse, proprietário deste conjunto dados e disponibilizado no formato CSV (*Comma-separated values*), separados por vírgula.

3.1.3 Estudo de caso

O estudo de caso é referente a informações contidas em letras de músicas e a intenção emocional existente nas palavras que compõem seu conteúdo, composto por

duas análises que utilizam o mesmo conjunto de dados, diferenciando-se apenas pelo objeto de interesse.

A primeira análise tem como objeto de interesse os termos unigramas, ou seja, uma simples palavra. A segunda tem como objeto de interesse os termos bigramas, ou seja, um par de palavras ou uma sequência de dois elementos adjacentes.

3.1.4 *Software* R Studio

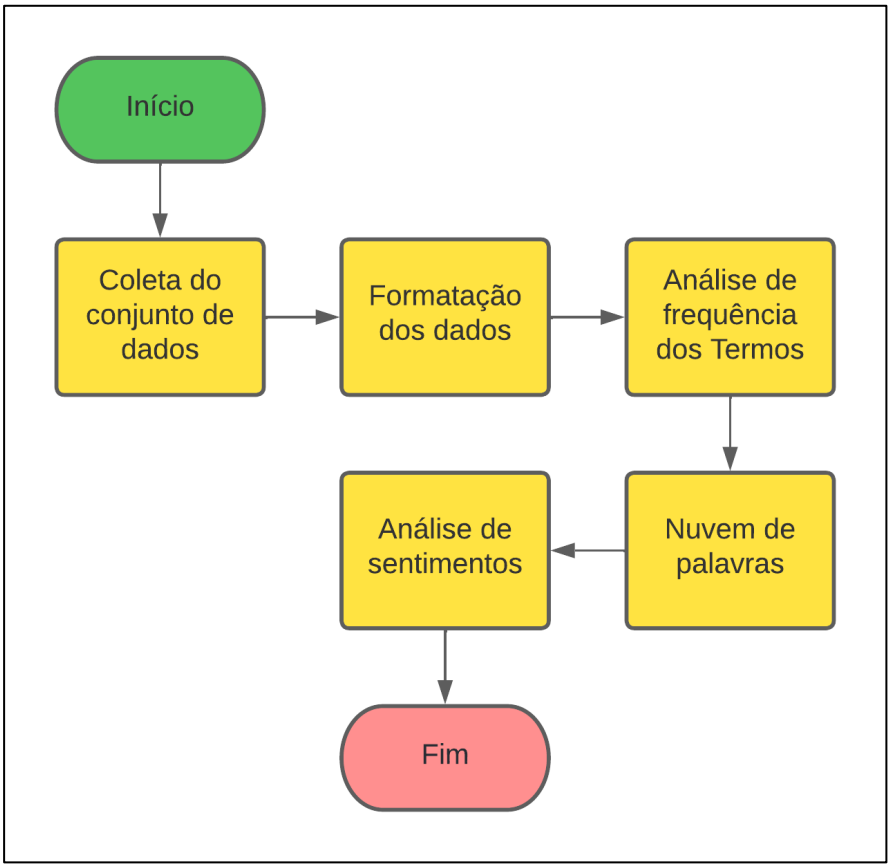
O *Software* utilizado na análise de dados neste trabalho foi o R Studio na versão 4.1.3, sendo escolhida por ser *software* gratuito e que contém vários recursos direcionados a análise de *Text Mining*, além de utilizar a linguagem de programação R que possui um amplo conjunto de pacotes e funções acessíveis para aplicações em diversas áreas do conhecimento, como: Ciência de dados (*Data Science*), Aprendizado de máquina (*Machine Learning*) e análise estatísticas. Ele se encontra disponível para download no endereço: https://vps.fmvz.usp.br/CRAN/ para todas as plataformas. A IDE (*Integrated Development Environment*) R Studio se encontra disponível para download no endereço: https://www.rstudio.com/products/rstudio/download/#download.

3.2 Métodos

A referência para conduzir as atividades propostas neste projeto de pesquisa inclui métodos e técnicas disponíveis na literatura voltada a pesquisa científica em computação e ciência de dados. O projeto é classificado de acordo com seus objetivos de natureza exploratório e experimental com abordagem predominantemente qualitativa, recorrendo a análises quantitativas utilizado do método de estudo de caso.

Para realização deste trabalho, foram definidas etapas que envolviam deste obter o conjunto de dados contendo as letras de músicas até extração e classificação de sentimentos envolvidos nas palavras que compunham as letras.

Figura 11- Fluxograma da abordagem proposta



Fonte: Elaborado pelo autor

A Figura 11 representa o fluxo de execução do método utilizado e as etapas necessários para o planejamento deste trabalho.

As letras de músicas foram extraídas do site brasileiro Vagalume e disponibilizadas pelo seu proprietário, um arquivo no formato CSV. De seu conteúdo original foi selecionado de forma manual, apenas o conteúdo suficientemente relevante para estre trabalho. Depois de selecionado, esses dados foram armazenados em objeto estruturado chamado: Corpus, que é formatado removendo o conteúdo não relevante para as análises. Através deste objeto formatado é possível aplicar as técnicas de mineração de textos, como: realizar a análise de frequência dos termos, gerar a nuvem de palavras e realizar a análise de sentimentos.

3.2.1 Análise de frequência dos termos

A análise de frequência dos termos disponibiliza informações contidas no conjunto de dados, possibilitando avaliar se a análise está seguindo o caminho correto, indicando os objetos de interesse supostamente esperados.

Na formatação dos dados ocorre a eliminação de informações que não são relevantes para a análise e podem até influenciar de forma errada, ou não desejada, como: números, pontos, acentos, espaços em branco e preposições que não agregam sentido específico na frase, tendo como objetivo ligar uma palavra a outra, e por isso não devem ser consideradas na análise. Essas palavras são chamadas de *stopwords*.

O conjunto de dados é transformado em uma matriz de termos do documento, ou matriz de termos frequentes, contendo as palavras em cada linha e os documentos (letras de músicas) em cada coluna. Sendo assim é possível os classificar e transformá-los em um *data frame*, contendo as palavras e quantidade de vezes que elas aparecem no documento.

3.2.2 Nuvem de palavras

Na nuvem de palavras é possível visualizar as informações contidas na análise de frequências dos termos, porém de uma forma gráfica e hierárquica em que quanto maior a frequência da palavra no conjunto de dados, maior é seu tamanho e mais ao centro a mesma estará.

3.2.3 Análise de sentimentos

Na análise de sentimentos é utilizado o método NRC do pacote *syuzhet* para categorizar a intenção emocional das palavras contidas no conjunto de dados após terem sido formatados. Neste processo, essas palavras são armazenadas em um vetor de caracteres que é submetido a análise de sentimentos de acordo com o dicionário do NRC que estabelece de forma binária ("sim” / “não”) para sentimentos positivos, negativos, ou se apresenta uma característica de outra emoção mais sutil, como: raiva, antecipação, nojo, medo, alegria, tristeza, surpresa ou confiança.

Depois de realizada essa categorização de cada palavra é possível identificar através de um gráfico de barras, o nível de cada sentimento em todo o conjunto de dados.

4 RESULTADOS

Este capítulo descreve os resultados dos experimentos realizados com o método proposto, apresentando os indicadores obtidos a partir das análises realizadas.

4.1 Conjunto de dados

O conjunto de dados utilizado para análise deste trabalho contém: 379.893 letras de músicas e 4.239 artistas.

A Tabela 1 representa as informações contidas neste conjunto de dados original. A coluna "Link para o perfil do artista no site: Vagalume" contém o perfil de cada artista ligado, a cada música coletada que em alguns casos se repetem pelo artista compor várias músicas. A coluna "Nome da música" contém o nome da música coletada, porém algumas músicas podem ter o mesmo nome e pertencer a artistas diferentes. A coluna "Link para a letra da música no site: Vagalume" contém o link exclusivo de acesso para cada música. A coluna "Letra da música" contém o conteúdo que compõe as letras das músicas, porém existem músicas que se repetem no conjunto de dados. A coluna "Idioma" contém o idioma em que a música foi escrita.

Tabela 1- Informações gerais da base de dados

Conjunto de dados – Original				
Link para o perfil do artista no site: Vagalume	**Nome da música**	**Link para a letra da música no site: Vagalume**	**Letra da música**	**Idioma**
4239	267259	379893	371182	50% inglês 41% português 9% outros

Fonte: Elaborador pelo autor

Por se tratar de um arquivo muito extenso, foi utilizado apenas uma parte deste conteúdo nos experimentos, mantendo de seu conteúdo original apenas 8.000 letras de músicas, em português no formado CSV. Todas as demais colunas e informações

foram removidas, por serem irrelevantes para análise em questão. A Tabela 2 indica as informações que foram removidas e o conteúdo mantido.

Tabela 2- Informações utilizadas da base de dados

Conjunto de dados – Utilizado				
Link para o perfil do artista no site: Vagalume	**Nome da música**	**Link para a letra da música no site: Vagalume**	**Letra da música**	**Idioma**
Informações removidas	Informações removidas	Informações removidas	8000	Informações removidas

Fonte: Elaborado pelo autor

4.2 Formatação dos dados

O conjunto de dados foi importado e armazenado em um objeto chamado: Corpus no *software* R. Esse objeto é utilizado quando se tem um conjunto grande e estruturado de texto, cujo objetivo seja realizar mineração.

Figura 12 - Formatação dos dados

```
# Limpando o objeto Corpus
# Converte o texto em minúsculo
corpus <- tm_map(corpus, content_transformer(tolower))
# Remove os números
corpus <- tm_map(corpus, removeNumbers)
# Remove as stopwords
corpus <- tm_map(corpus, removeWords, stopwords('portuguese'))
# Remove os pontos
corpus <- tm_map(corpus, removePunctuation)
# Remove os espaços em branco
corpus <- tm_map(corpus, stripWhitespace)
```

Fonte: Elaborado pelo autor

Foi utilizado o pacote *Text Mining* (TM) fornecido pela linguagem R para a formatação necessária dos dados. Esse pacote fornece diversos métodos para limpeza dos dados, conforme representado na Figura 12, tais como: converter o texto

em minúsculo, remover os números, as *stopwords*, os pontos, os espaços em branco entre outros.

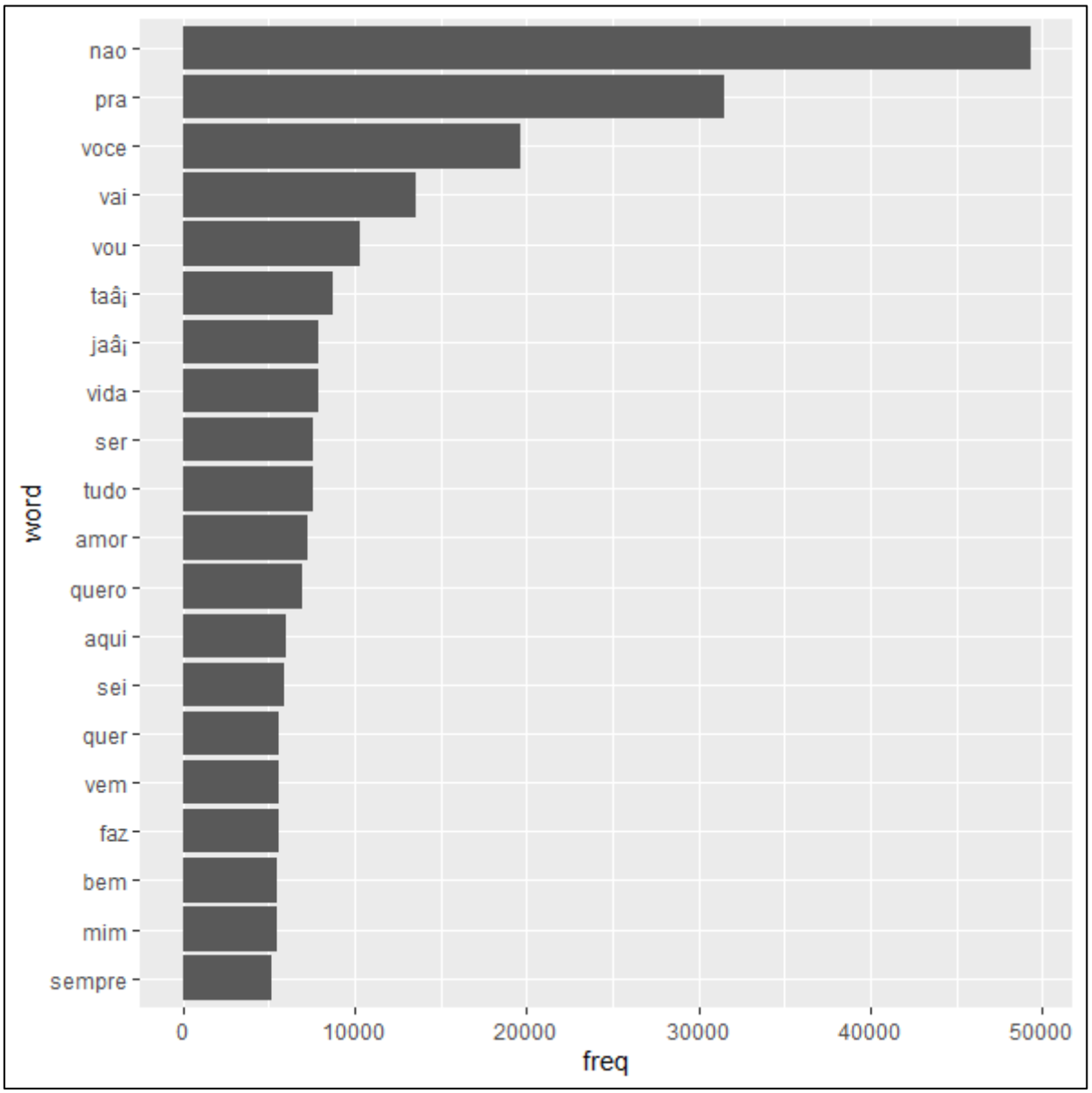
4.3 Análise dos termos unigramas

Neste item são apresentadas as técnicas de mineração de texto baseadas em unigramas, ou seja, palavras simples, aplicadas ao conjunto de dados, e o resultados gerados a partir delas. Os resultados obtidos se baseiam nas análises de frequência dos termos, nuvem de palavras e análise de sentimentos.

4.3.1 Análise da frequência termos

Para que seja possível realizar a análise da frequência de cada palavra contida nas letras das músicas selecionadas é necessário criar um objeto chamado matriz de termos do documento, ou matriz de termos frequentes. Esse objeto é criado através do método "*TermDocumentMatrix (corpus)*" que tem como parâmetro o objeto Corpus após a sua limpeza e formatação dos dados. Essa matriz é classificada e convertida em um *data frame* que contém as palavras e a frequência em que aparecem, utilizado para realizar as análises.

Figura 13 - Frequência de termos unigrama



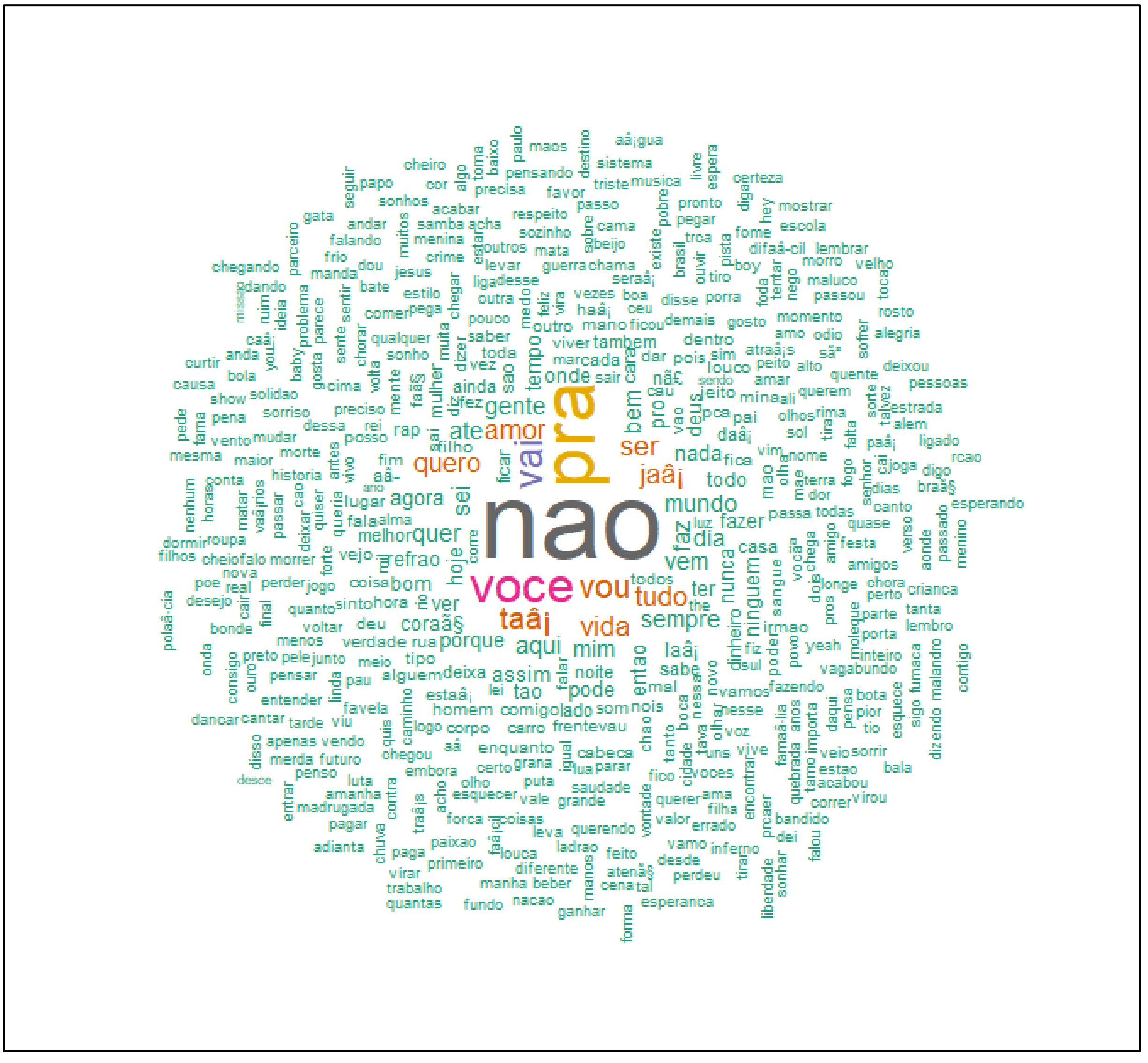
Fonte: Elaborado pelo autor

Na Figura 13 é possível identificar as 20 palavras mais frequentes, tendo a palavra “nao” como a que possui a maior frequência dentro conjunto de dados. Existem palavras que se encontram desconfiguradas após importação, como por exemplo: “taâ¡”, “jaâ¡” entre outras.

4.3.2 Nuvem de palavras

A Figura 14 representa a nuvem de palavras gerada baseada no *data frame* informado no item 4.3.1 através do pacote *wordcloud*, contendo as 500 palavras de maior frequência dentro do conjunto de dados. Quanto maior a sua frequência, maior o tamanho de sua fonte e mais ao centro estará.

Figura 14- Nuvem de palavras unigrama (*wordcloud*)



Fonte: Elaborado pelo autor

A Figura 15 representa a nuvem de palavras gerada baseada no *data frame* informado no item 4.3.1 através do pacote *wordcloud2*, contendo “todas” as palavras

do conjunto de dados de acordo com sua frequência. Quanto maior a sua frequência, maior o tamanho de sua fonte e mais ao centro estará.

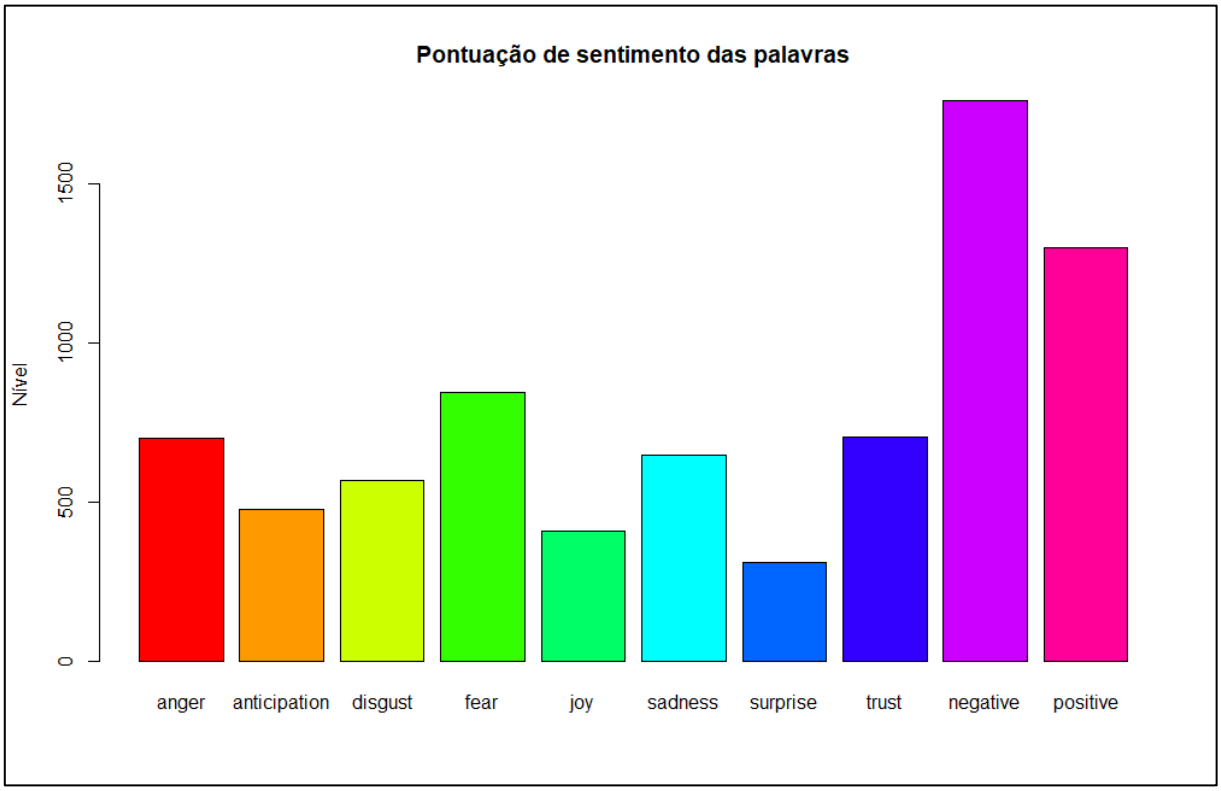
Figura 15 - Nuvem de palavras unigrama (*wordcloud2*)

Fonte: Elaborado pelo autor

4.3.3 Análise de sentimento

O gráfico ilustrado na Figura 16 representa a utilização do método NRC para categorizar e estabelecer o nível de intenção emocional mais sutil das palavras contidas no conjunto de dados, como: raiva, antecipação, nojo, medo, alegria, tristeza, surpresa, confiança, além dos sentimentos positivos e negativos.

Figura 16 - Nível de sentimento das palavras unigrama



Fonte: Elaborado pelo autor

4.4 Análise dos termos bigrama

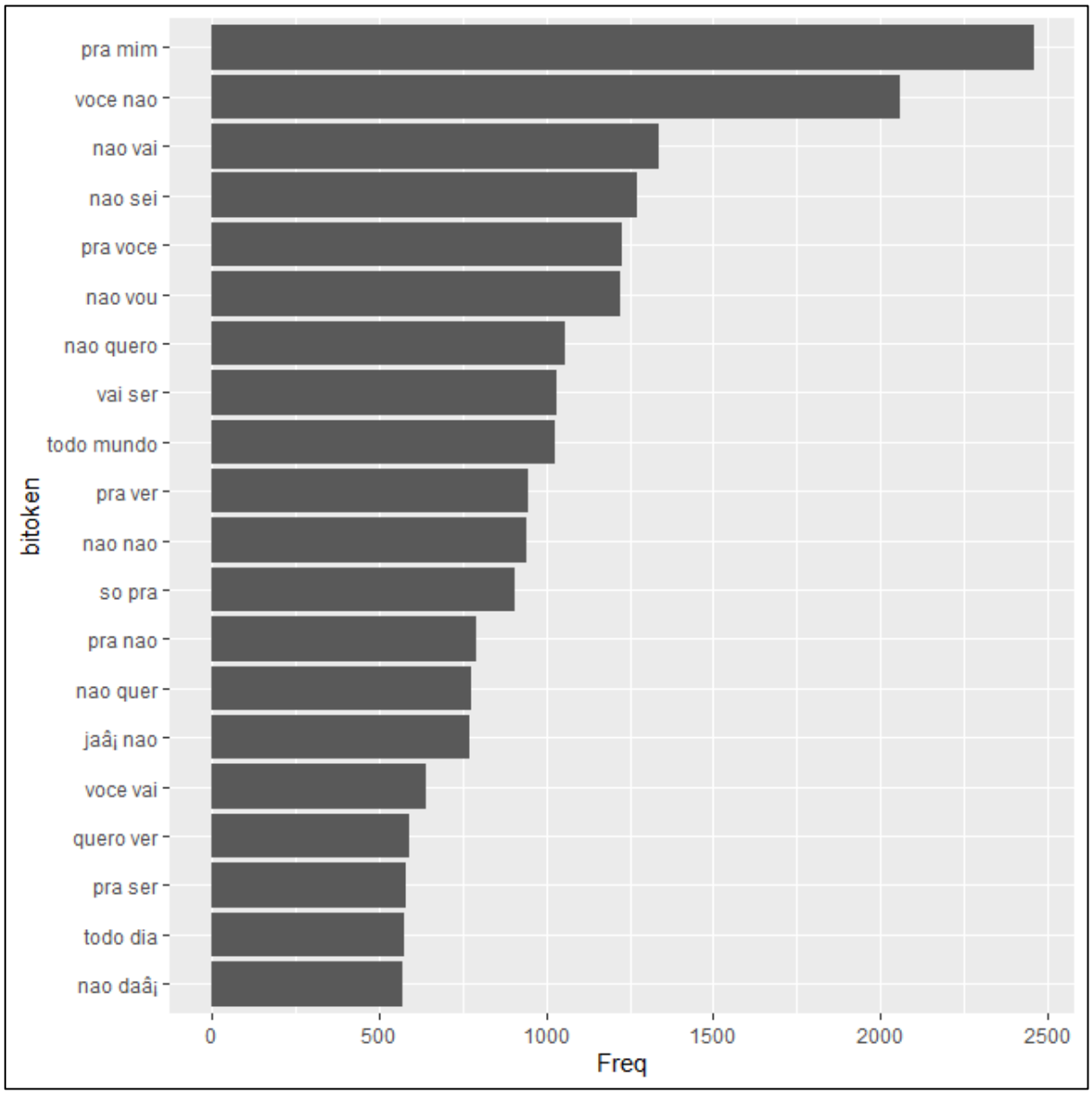
Neste item são apresentadas as técnicas de mineração de texto baseadas em correlação entre palavras, examinando quais palavras tendem a seguir outras imediatamente. Este processo se baseia na relação bigrama, ou seja, um par de palavras, ou uma sequência de dois elementos adjacentes, aplicadas ao conjunto de dados, e o resultados gerados a partir delas. Os resultados obtidos se baseiam nas análises de frequência dos termos adjacentes, nuvem de palavras e análise de sentimentos.

4.4.1 Análise da frequência termos

Para que seja possível realizar a análise da frequência de cada conjunto de palavras adjacentes contida nas letras das músicas selecionadas é necessário definir o valor mínimo dessa relação, neste caso, definindo o **n** como sendo igual a **2**. Feito isso, é criado um delimitador para pode separar esses conjuntos de palavras de interesse das demais.

O processo de tokenização é realizado pelo pacote RWeka, utilizando o método ”*NGramTokenizer(corpus, weka_control(min, max, delimiters))*” que tem como parâmetro o objeto Corpus após usa limpeza e formatação dos dados, além do tamanho mínimo e máximo dos *tokens*, e o delimitador criado para os separar. Esse objeto é classificado e convertido em um *data frame* que contém o as palavras adjacentes e a frequência em que aparecem juntas no conjunto de dados, utilizado para realizas as análises.

Figura 17 - Frequência de termos bigrama



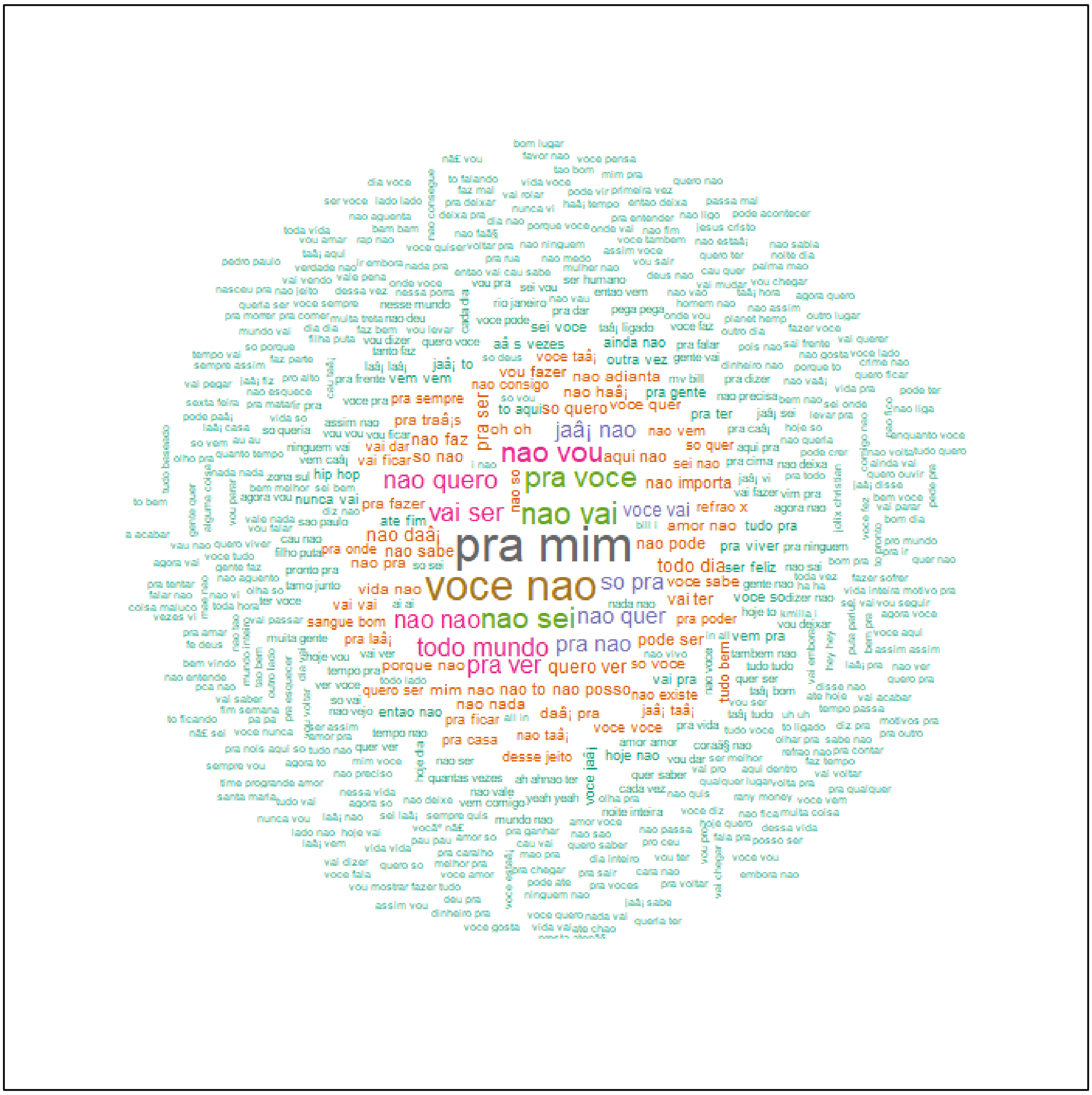
Fonte: Elaborado pelo autor

Na Figura 17 é possível identificar as 20 palavras adjacentes mais frequentes, tendo as palavras “pra mim” como as que possuem a maior frequência juntas dentro conjunto de dados, permitindo visualizar que palavras tendem a ocorrer imediatamente, possibilitando até mesmo uma predição de ocorrência. Existem palavras que se encontram desconfiguradas após importação, como por exemplo: “jaâ¡ nao”, “nao daâ¡” entre outras.

4.4.2 Nuvem de palavras

A Figura 18 representa a nuvem de palavras gerada baseada no *data frame* informado no item 4.4.1 através do pacote *wordcloud*, contendo as 500 palavras adjacentes de maior frequência dentro do conjunto de dados. Quanto maior a sua frequência, maior o tamanho de sua fonte e mais ao centro estará.

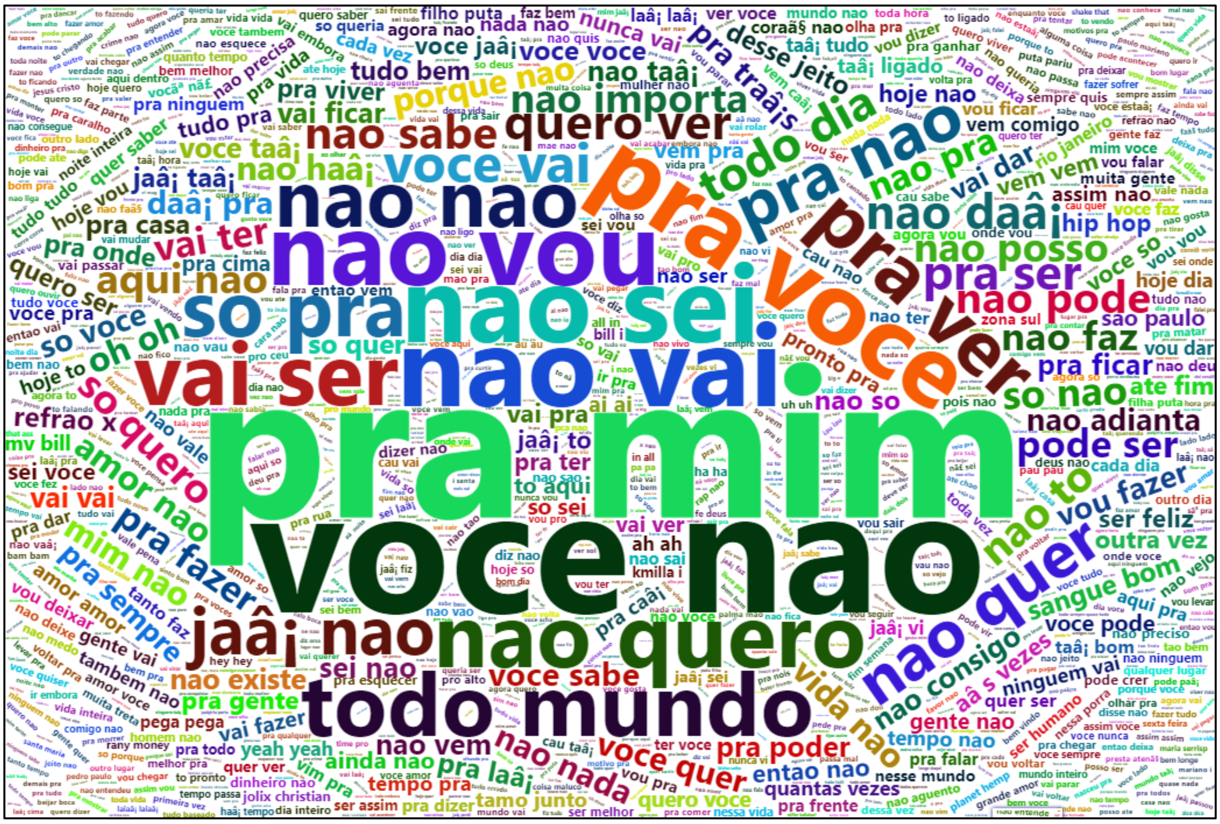
Figura 18- Nuvem de palavras bigrama (*wordcloud*)



Fonte: Elaborado pelo autor

A Figura 19 representa a nuvem de palavras gerada baseada no *data frame* informado no item 4.4.1 através do pacote *wordcloud2*, contendo “todas” as palavras adjacentes do conjunto de dados de acordo com sua frequência. Quanto maior a sua frequência, maior o tamanho de sua fonte e mais ao centro estará.

Figura 19 - Nuvem de palavras bigrama (*wordcloud2*)

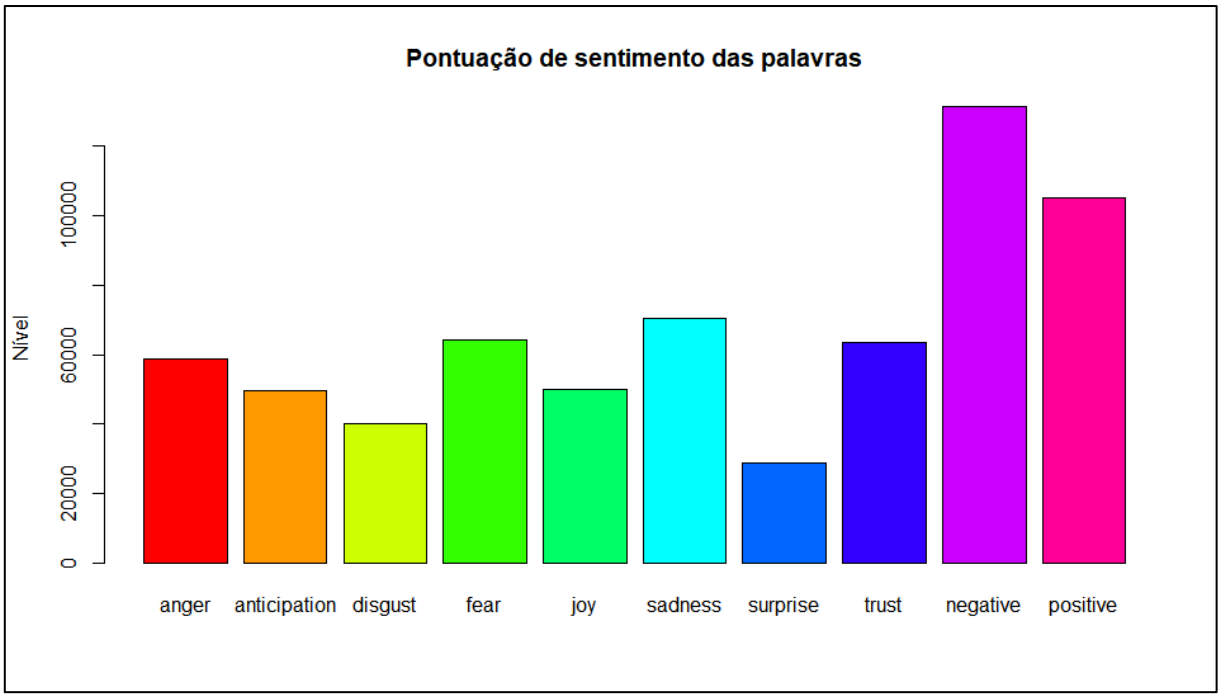


Fonte: Elaborado pelo autor

4.4.3 Análise de sentimento

O gráfico ilustrado na Figura 20 representa a utilização do método NRC para categorizar e estabelecer o nível de intenção emocional mais sutil das palavras adjacentes contidas no conjunto de dados, como: raiva, antecipação, nojo, medo, alegria, tristeza, surpresa, confiança, além dos sentimentos positivos e negativos.

Figura 20 - Nível de sentimento das palavras bigrama



Fonte: Elaborado pelo autor

5. CONCLUSÃO

Este trabalho apresentou o processo de descoberta de conhecimento de um conjunto dados aplicados a letras de músicas, detalhando o processo de limpeza e formatação destes dados e a aplicação das técnicas de mineração de textos, como: frequência de termos, nuvem de palavras, análise de sentimento e correlação entre palavras.

O *software* R disponibiliza diversos pacotes para realização de mineração de textos de forma gratuita e de fácil acesso. Durante a realização deste trabalho foram encontrados alguns problemas na utilização deste conjunto de dados, como por exemplo: o conjunto de dados é bastante extenso, sendo preciso descartar boa parte do seu conteúdo para ser possível a realização dos experimentos, e ainda assim, diversas manipulações demoravam até mais de 30 minutos para serem executadas. Realizado a revisão ortográfica das letras musicais com auxílio do Excel, o conjunto de dados também apresentou diversas palavras desconfiguradas e com erros ortográficos, principalmente após sua importação no R. Durante o processo de análise de sentimentos, algumas palavras não continham intenção emocional devido não serem encontradas no dicionário do método utilizado, por estarem desconfiguras ou apresentarem erros ortográficos.

Apesar dos problemas e dificuldades encontrados durante a realização dos experimentos neste trabalho, os resultados foram satisfatórios e o modelo de classificação gerado neste trabalho permite identificar o conteúdo e o sentimento envolvido no conjunto de dados utilizado de modo a observar sentimentos negativos e egocêntricos, presentes em termos de maior frequência, tais como: não quero, não vou, não sei, não vai, pra mim entre outros.

Na busca de obter melhores resultados, é possível utilizar a API disponibilizada pelo Spotify que possui uma quantidade muito interessante de informações sobre música e álbuns, acessando os dados da API (*Application Programming Interface*) pelo próprio R através da biblioteca spotifyR, além de realizar a própria coleta das letras de músicas para se montar um conjunto de dados mais consistente.

Também é possível utilizar a biblioteca GoogleLanguageR para realizar a tradução dos textos e para obter resultados mais concretos quanto a análise de sentimentos, já que os dicionários dos métodos se encontram formatados por palavras em inglês, utilizar outros métodos de análise de sentimento como AFFIN (Finn Ärup

Nielsen) e BING (Bing Liu), comparando seus resultados para o mesmo conjunto de dados. Por fim, é possível criar um dicionário de palavras de interesse para consultar no conjunto de dados se elas aparecerem e realizar uma filtragem delas para determinada tratativa.

REFERÊNCIAS

AIDAR, Laura**;** **História da Música.** Disponível em: <https://www.todamateria.com.br/historia-da-musica/> Acesso em: 24 maio 2022

BERRY, M. W.; KOGAN, J. **Text mining: applications and theory**. [S.l.]: John Wiley & Sons, 2010.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. **Mineração de dados: Conceitos, tarefas, métodos e ferramentas**. Universidade Federal de Goiás (UFC), p. 1-29, 2009.

CHOWDHURY, G. G. **Natural language processing**. *Annual review of information science and technology*, Wiley Online Library, v. 37, n. 1, p. 51–89, 2003.

Compreendendo a Música - Disciplina – Arte. Disponível em: <http://www.arte.seed.pr.gov.br/modules/conteudo/conteudo.php?conteudo=136> Acesso em: 24 maio 2022.

ESTEVES, Wilson; **O que é a música?**. Disponível em: <https://www.musicdot.com.br/artigos/saiba-o-que-e-musica?gclid=Cj0KCQjwhLKUBhDiARIsAMaTLnG2bOuyoy-eJzG9LmXvV91nmw0GgJ8ptC9vUlb132ZLnc1RtPMfEYwaAuRyEALw_wcB>. Acesso em: 24 maio 2022.

FAYYAD, U.; PIATESKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. **Advances in knowledge discovery and data mining**. Cambridge: MIT Press, 1996. 560p.

FEINERER, I.; HORNIK, K; MEYER, D. **Text Mining Infrastructure in R**. Journal of Statistical Software, 2008.

FREITAS, Henrique Mello Rodrigues de; JANISSEK-MUNIZ, Raquel; MOSCAROLA, Jean. **Uso da Internet no processo de pesquisa e análise de dados.** Associação Nacional de Empresas de Pesquisa (2004: São Paulo).

FREITAS, João Vitor; **Mineração de Textos: Análise de Sentimentos no Twitter Referentes às Eleições para Presidente do Brasil em 2018**. Pontifícia Universidade Católica de Goiás (PUC- GO), 2018.

FREITAS, L. d; VIEIRA, R. **Exploring resources for sentiment analysis in portuguese language**. In: IEEE. 2015 Brazilian Conference on Intelligent Systems (BRACIS). [S.l.], 2015. p. 152–156.

Gêneros musicais e suas definições. Disponível em: <https://www.collectorsroom.com.br/2013/09/generos-musicais-e-suas-definicoes.html?m=1>. Acesso em: 24 maio 2022.

JUNIOR, Jorge L. F. S; ROSSI, Rafael G.; LOBATO Fábio M. F.; **A Lyric-Based Approach for Brazilian Music KnowledgeDiscovery: Brazilian Country Music as a Case Study**, Universidade Federal de Mato Grosso do Sul - Três Lagoas, 2019.

KWARTLER, Ted. **Text mining in practice with R**. John Wiley & Sons, 2017.

LIMA, Higor Gomes; **Estudo de Caso para Extração de Informação baseado em Ontologia**. Pontifícia Universidade Católica de Goiás (PUC- GO), 2019.

MARQUESONE, Rosangela. **Big Data: técnicas e tecnologias para extração de valor dos dados**. Editora Casa do Código, 2017.

Música? | MusicDot. Disponível em: <https://www.musicdot.com.br/artigos/saiba-o-que-e-musica?gclid=Cj0KCQjwhLKUBhDiARIsAMaTLnG2bOuyoy-eJzG9LmXvV91nmw0GgJ8ptC9vUlb132ZLnc1RtPMfEYwaAuRyEALw_wcB>. Acesso em: 24 maio 2022.

SILGE, Julia; ROBINSON, David. **Text mining with R: A tidy approach**. “O'Reilly

Media, Inc.", 2017.

Song lyrics from 79 musical genres. Disponível em: <https://www.kaggle.com/neisse/scrapped-lyrics-from-6-genres?select=lyrics-data.csv>. Acesso em: 15 out 2021.

SOUZA, Renato Rocha; CAFE, Lígia Maria Arruda; **Análise de Sentimento Aplicada ao Estudo de Letras de Música**, 2018.

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
GABINETE DO REITOR
Av. Universitária, 1069 • Setor Universitário
Caixa Postal 86 • CEP 74605-010
Goiânia • Goiás • Brasil
Fone: (62) 3946.1000
www.pucgoias.edu.br • reitoria@pucgoias.edu.br

RESOLUÇÃO nº 038/2020 – CEPE

ANEXO I

APÊNDICE ao TCC

Termo de autorização de publicação de produção acadêmica

O(A) estudante Igor Ferreira de Jesus Pereira do Curso de Ciência da Computação, matrícula 2015.2.0028.0022-3, telefone: 62 992044261 e-mail igorferreiralian@gmail.com, na qualidade de titular dos direitos autorais, em consonância com a Lei nº 9.610/98 (Lei dos Direitos do Autor), autoriza a Pontifícia Universidade Católica de Goiás (PUC Goiás) a disponibilizar o Trabalho de Conclusão de Curso intitulado Mineração de texto aplicada à identificação de sentimentos e intenções, gratuitamente, sem ressarcimento dos direitos autorais, por 5 (cinco) anos, conforme permissões do documento, em meio eletrônico, na rede mundial de computadores, no formato especificado (Texto(PDF); Imagem (GIF ou JPEG); Som (WAVE, MPEG, AIFF, SND); Vídeo (MPEG, MWV, AVI, QT); outros, específicos da área; para fins de leitura e/ou impressão pela internet, a título de divulgação da produção científica gerada nos cursos de graduação da PUC Goiás.

Goiânia, 22 de junho de 2022.

Assinatura do autor: Igor Ferreira de Jesus Pereira

Nome completo do autor: Igor Ferreira de Jesus Pereira

Assinatura do professor–orientador: Sibelius Lellis Vieira

Nome completo do professor-orientador: Sibelius Lellis Vieira