UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO

Julio Pancracio Valim

Classificação de sentimento em arquivos de áudio a partir da extração de Coeficientes Cepstrais na frequência Mel (MFCCs)

São Carlos

2021

Julio Pancracio Valim

Classificação de sentimento em arquivos de áudio a partir da extração de Coeficientes Cepstrais na frequência Mel (MFCCs)

Trabalho de conclusão de curso apresentado ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, como parte dos requisitos para conclusão do MBA em Ciências de Dados.

Área de concentração: Ciências de Dados

Orientador: Prof. Dr. Marcelo Garcia Manzato

São Carlos

2021

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

S856m

Valim, Julio Pancracio
Classificação de sentimento em arquivos de áudio a partir da extração de Coeficientes Cepstrais na frequência Mel (MFCCs) / Julio Pancracio Valim ; orientador Marcelo Garcia Manzato. – São Carlos, 2021.
17 p. : il. ; 30 cm.

Monografia (MBA em Ciências de Dados) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2021.

1. Ciência de Dados. 2. Psicoacústica. 3. Classificação de sentimentos. I. Manzato, Marcelo Garcia, orient. II. Título.

Julio Pancracio Valim

Classificação de sentimento em arquivos de áudio a partir da extração de Coeficientes Cepstrais na frequência Mel (MFCCs)

Trabalho de conclusão de curso apresentado ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, como parte dos requisitos para conclusão do MBA em Ciências de Dados.

Data de defesa: 22 de janeiro de 2022

Comissão Julgadora:

Prof. Dr. Marcelo Garcia Manzato
Orientador

Professor
Convidado1

Professor
Convidado2

São Carlos
2021

RESUMO

VALIM, J. P. **Classificação de sentimento em arquivos de áudio a partir da extração de Coeficientes Cepstrais na frequência Mel (MFCCs)**. 2021. 17p. Monografia (MBA em Ciências de Dados) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2021.

Resumo em desenvolvimento.

Esta pesquisa investiga o desempenho comparativo entre modelos de classificacão multiclasses na categorização de arquivos de áudio segundo o sentimento hegemônico, isto é, positivo, negativo ou neutro. Para tanto, realizou-se análise comparativa entre o desempenho obtido com os classificadores *SVM* e *CNN*, treinados com dados coletados a partir da extração dos Coeficientes Cepstrais na frequência Mel (MFCCs) de segmentos de áudio. A base de dados para desenvolvimento da pesquisa foi elaborada a partir de áudios selecionados e pré categorizados. Por resultado, observou-se ...

Ciência de Dados; Psicoacústica; Classificação de Sentimentos;:

ABSTRACT

VALIM, J. P. **Sentiment classification in audio files based on Cepstral Coefficients in the Honey frequency (MFCCs)**. 2021. 17p. Monografia (MBA em Ciências de Dados) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2021.

In progress.

This research works on a comparison between performance from multi class classification models for audio files categorization according to hegemonic sentiment, which are positive, negative or neutral. Therefore, an analysis was performed between results obtained with the classifiers *SVM* and *CNN*, trained with data obtained from the extraction of Melody Frequency Cepstral Coefficients (MFCCs) of audio segments. The database for research development was created from selected and precategorized audios.

Data Science; Psychoacoustic; Sentiment Analysis;

LISTA DE ABREVIATURAS E SIGLAS

CNN	Convolutional Neural Network
HMM	Hidden Markov Model
LFPC	Log Frequency Power Coefficients
LSTM	Long-Short Term Memory Neural Network
MFCC	Mel-Frequency Cepstral Coefficients
PNL	Processamento de Linguagem Natural
SVM	Support Vector Machine Classifier

SUMÁRIO

1 INTRODUÇÃO

A curiosidade humana movimenta, entre variadas iniciativas, também a de superar limitações físicas, temporais e cognitivas. A incipiente Ciências da Computação, aderiu também ao desafio no momento em que TURING (1950) estabeleceu a pedra angular das pesquisas em inteligência artifical: é possível uma máquina imitar o comportamento humano? Por certo a questão não tem resposta simples ou definitiva, haja vista os desdobramentos éticos e filosóficos do tema, porém, sete décadas depois, presenciamos avanços significativos e surpreendentes na condução automatizada de veículos, nos exames diagnósticos e intervenções cirúrgicas de saúde, nas tendências de consumo e projeções de preços, nas previsões meteorológicas, na otimização do ensino, nos atendimentos e diálogos diversos entre dispositivos digitais e seres humanos, entre outras aplicações.

Em relação ao tópico conversacional, uma vertente das ciências da computação associada aos estudos linguísticos desenvolveu a área de processamento de linguagem natural (PLN), responsável pelas soluções que capacitam computadores à compreensão de signos textuais e produção de respostas em formato inteligível a qualquer pessoa versada no idioma utilizado. A operação de códigos no nível da linguagem humana comum proporciona aplicabilidade do processamento digital a uma ampla gama de problemas e cria a possibilidade de leitura de enormes bases de textos. Nesse âmbito, a classificação de trechos de escrita segundo o sentimento dominante é um exemplo relevante dos avanços da área, largamente utilizada no intuito de conhecer em mais detalhes uma base de usuários e implementar medidas que melhorem a experiência da interação humano-computador. Entre ganhos oriundos da análise de sentimento destaca-se, como exemplos, a medição da satisfação do cliente e recomendação de produtos, amplamente utilizada no marketing e serviços de atendimento, o mapeamento de opiniões políticas e intenções de voto, totalmente incorporados à decisões de governos, partidos políticos e instituições públicas, e a tendência em relação à negócios da área financeira, com o objetivo de mensurar liquidez de mercado, prever movimentos de alta ou queda em preços de ações e identificar fatores de risco em mercados futuros.

Na atividade de detecção de sentimento em textos, a apuração e tratamento da escrita orientam-se de modo apropriado às características desse tipo de dado, lidando, por exemplo, com contagem de caracteres ou de tokens, semelhança entre frases e detecção de palavras chaves ou entidades, além de tratamentos morfológicos de estemização e lematização, os quais resultam na identificação do radical das palavras e da forma infinitiva dos verbos, respectivamente. Embora hajam variados procedimentos, melhor adequados a estrutura dos dados processados, sejam eles textuais, visuais ou sonoros, é possível, em suma, generalizar o escopo das análises de sentimento como restrito a três finalidades

principais: identificar polaridades, isto é, positiva, negativa, ou neutra, reconhecer classes mais específicas em relação às polaridades, ou seja, alegria, tristeza, esperança, frustração, medo, etc., e, de maneira mais específica, mapear opiniões em relação a um tema, por exemplo, o que pensam os jovens brasileiros em relação à ciência – aprovam/desaprovam, confiam/desconfiam e assim por diante.

A despeito de todo êxito alcançado e a franca expansão da pesquisa nesse domínio, algumas características ainda dificultam o desenvolvimento da análise de sentimento sobre dados textuais. É possível listar ao menos dois fatores preponderantes, o idioma, com seu alfabeto, sintaxe, morfologia, regras gramaticais e demais estruturas específicas, e além dele os elementos que apoiam a funcionalidade da linguagem, isto é, o contexto em seus múltiplos aspectos culturais, sociais, etários, comerciais, geográficos, etc.

Como alternativa ao dado textual para o propósito de classificação de sentimentos, um campo bastante consolidado das ciências físicas, a psicoacústica, fornece elementos basilares de uma abordagem promissora e amplia o campo de pesquisa em PNL. A identificação e categorização de sentimentos em documentos no formato de áudio é uma área relativamente nova e em experimentação que, contudo, apresenta resultados relevantes e avanços expressivos ano após ano, desde a introdutória abordagem de classificação por meio de um modelo de Markov (HMM) aplicado sobre descritores dos sinais de voz como coeficientes de potência de freqüência logarítmica (LFPC) gerados em janelas curtas de tempo, publicado por Nwe, Foo e De Silva (2003), passando pelo trabalho de Chew et al. (2011), com o desenvolvimento de classificadores paralelos de features sonoras como pitch e coeficiente cepstral fracionário, a pesquisa de Xie e Guan (2012) em abordagem referenciada por fusão de informação e descritor de entropia, ou ainda por estudos mais recentens como a proposta de transferência de conhecimento obtido por reconhecimento facial para o áudio associado por meio de redes generativas semi-supervisionadas, desenvolvida por He et al. (2020).

Na esteira destas pesquisas, portanto, o estudo exposto nas páginas subsequentes propõe-se a investigar a aplicabilidade de um modelo de aprendizagem para categorização de arquivos de áudio segundo o sentimento hegemônico observado, isto é, positivo, negativo ou neutro. De modo mais específico, a partir de dados extraídos de segmentos de áudio, ou seja, diferentes combinações entre rate, frequência e interpolação intervalar derivados da representação cepstral melódica do sinais sonoros (MFCCs), obtidas pela abordagem não paramétrica no domínio da frequência, foram testados os modelos SVM e CNN, selecionados por meio de revisão bibliográfica, e realizada comparação crítica dos resultados.

2 PANORAMA

Após duas décadas de progressos nas metodologias classificação de sentimentos em áudio, variadas formatações de dados e técnicas de modelagem foram experimentadas. Resumindo de modo didático desde conceitos básicos até o atual estado da arte com o emprego de redes neurais híbridas, Zhao, Ye e Wang (2018) apresentam elucidativo panorama dessas abordagens.

Parte significativa do desafio da classificação de sentimentos reside nos pressupostos teóricos do tema. De antemão, por tratar-se de um problema de classificação os resultados estão diretamente relacionados à definição *a priori* dos sentimentos, ou seja, os limites que definem cada classe. Dessa forma é válido ressaltar que a definição prévia do domínio dos sentimentos está diretamente relacionada aos resultados obtidos. Do ponto de vista técnico, tais resultado variam de acordo com a qualidade do dado de *input*, se ele é usado no formato original, convertido para espectograma, ou ainda transformado em conversão de escala ou de frequência.

Logo, uma característica fundamental do processamento de áudio, é a natureza não estacionária desse tipo de dado, a qual permite a extração de atributos no domínio temporal ou da frequência para dados originais, através do cálculo de descritores como:

1. Cruzamentos por zero: taxa de mudanças de sinal (positivo/negativo) durante uma janela temporal;
2. Energia: soma dos quadrados das amplitudes normalizada pelo tamanho da janela temporal;
3. Energia da entropia: entropia das energias normalizadas das janelas temporais (medida relaciona à mudanças abruptas);
4. Momentos estatísticos: cálculos estatísticos da média, desvio padrão, obliquidade, curtose, etc, do sinal global.

A produção de descritores é igualmente válida para representações espectrais de trechos de áudio. Para casos de aplicação da transformada de Fourier, é possível gerar descritores do centro espectral, da dispersão do espectro e da entropia espectral, isto é, a entropia das energias espectrais normalizadas, para um conjunto de sub-janelas temporais. Outra abordagem largamente empregada para produção de descritores a partir de trechos de som resulta da extração de coeficientes cepstrais da frequência melódica, os MFCCs. No processamento do som, o cepstrum representa o espectro de potência de um som para uma janela de curto prazo, obtida pela transformada de cosseno linear de um espectro de

potência logarítimo, de forma que sua sequência capta as propriedades acústicas de sinais de voz e resume a forma como o ouvido humano percebe o som da fala.

No que diz respeito ao processo da classificação dos sinais de áudio, é possível categorizá-los segundo o tipo da sua implementação. O procedimento nomeado tradicional procede segundo a sequência clássica das ciências de dados, a qual se inicia com o preprocessamento, seguido da extração de features, redução de dimensionalidade e seleção do classificador. Outra tendência filia-se à utilização de redes neurais como forma de diminuir a carga de conhecimento prévio necessário à execução do processo, valendo-se de modelos avançados para o treinamento e classificação implementados sobre os dados originais. Um modo intermediário orienta-se próximo ao processo tradicional, recorrendo, no entanto, ao apoio de redes neurais para realização das tarefas de extração de features e redução de dimensionalidade.

Capítulo em desenvolvimento.

3 METODOLOGIA

Em processo de elaboração conforme item 3 do cronograma de desenvolvimento deste trabalho de conclusão de curso.

1. *Levantamento bibliográfico e mapeamentos de técnicas de classificação de áudio.*
2. *Captação de coeficientes cepstrais e tratamento da base de dados.*
3. *Comparação entre as principais técnicas estudadas e análise de viabilidade via simulação.*
4. *Aplicação dos modelos em dados reais.*
5. *Discussão dos resultados obtidos e revisão do TCC.*

4 RESULTADOS

Em processo de elaboração conforme itens 4 e 5 do cronograma de desenvolvimento deste trabalho de conclusão de curso.

1. *Levantamento bibliográfico e mapeamentos de técnicas de classificação de áudio.*
2. *Captação de coeficientes cepstrais e tratamento da base de dados.*
3. *Comparação entre as principais técnicas estudadas e análise de viabilidade via simulação.*
4. *Aplicação dos modelos em dados reais.*
5. *Discussão dos resultados obtidos e revisão do TCC.*

5 CONCLUSÃO

Em processo de elaboração conforme item 5 do cronograma de desenvolvimento deste trabalho de conclusão de curso.

1. *Levantamento bibliográfico e mapeamentos de técnicas de classificação de áudio.*
2. *Captação de coeficientes cepstrais e tratamento da base de dados.*
3. *Comparação entre as principais técnicas estudadas e análise de viabilidade via simulação.*
4. *Aplicação dos modelos em dados reais.*
5. *Discussão dos resultados obtidos e revisão do TCC.*

REFERÊNCIAS

CHEW, L. W. et al. Audio-emotion recognition system using parallel classifiers and audio feature analyzer. In: **2011 Third International Conference on Computational Intelligence, Modelling Simulation**. [s.n.], 2011. p. 210–215. Disponível em: <https://ieeexplore.ieee.org/document/6076358>. Acesso em: 30 mai. 2021.

HE, G. et al. Image2audio: Facilitating semi-supervised audio emotion recognition with facial expression image. In: **2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)**. [s.n.], 2020. p. 3978–3983. Disponível em: <https://ieeexplore.ieee.org/document/9150723>. Acesso em: 10 jul. 2021.

NWE, T. L.; FOO, S. W.; De Silva, L. C. Speech emotion recognition using hidden markov models. **Speech Communication**, v. 41, n. 4, p. 603–623, 2003. ISSN 0167-6393. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167639303000992>. Acesso em: 4 jun. 2021.

TURING, A. M. I.—COMPUTING MACHINERY AND INTELLIGENCE. **Mind**, LIX, n. 236, p. 433–460, 10 1950. ISSN 0026-4423. Disponível em: <https://doi.org/10.1093/mind/LIX.236.433>. Acesso em: 9 jul. 2021.

XIE, Z.; GUAN, L. Multimodal information fusion of audio emotion recognition based on kernel entropy component analysis. In: . [s.n.], 2012. v. 07, p. 1–8. Disponível em: <https://ieeexplore.ieee.org/document/6424622>. Acesso em: 4 jun. 2021.

ZHAO, H.; YE, N.; WANG, R. A survey on automatic emotion recognition using audio big data and deep learning architectures. In: **2018 IEEE 4th International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing, (HPSC) and IEEE International Conference on Intelligent Data and Security (IDS)**. [s.n.], 2018. p. 139–142. Disponível em: <https://ieeexplore.ieee.org/document/8552297>. Acesso em: 10 jul. 2021.