UNIVERSIDADE FEEVALE

GUILHERME LAUTERT

SISTEMA DE REGISTRO E CONSOLIDAÇÕES DE GASTOS E FINANÇAS UTILIZANDO RECONHECIMENTO DE INSTRUÇÕES POR VOZ

Novo Hamburgo

2022

GUILHERME LAUTERT

SISTEMA DE REGISTRO E CONSOLIDAÇÕES DE GASTOS E FINANÇAS UTILIZANDO RECONHECIMENTO DE INSTRUÇÕES POR VOZ

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do grau de Bacharel em Ciência da computação pela Universidade Feevale

Orientador: Juliano Varella de Carvalho

Novo Hamburgo

2022

AGRADECIMENTOS

Gostaria de agradecer a minha esposa, dando força e apoio, durante a realização deste Trabalho de Conclusão de Curso.

Ao meu orientador, Juliano Varella, por todas das conversas que tivemos durante o curso de graduação, trocas de ideia, incentivos e em especial pelo auxilio no desenvolvimento deste TCC, criticando, apoiando e fornecendo o caminho a ser tomado.

Aos professores que tive durante a vida, com seus ensinamentos, auxiliando em minha caminhada e tomada de decisão. Em especial as professoras, Adriana Neves dos Reis, Sandra Teresinha Miorelli e ao professor Gabriel da Silva Simões, por todas conversações e trocas de ideia sobre programação, metodologia, padrões e gerencia de projetos, organização, etc. após as aulas.

Aos meus pais e irmã, que contribuíram com as minhas escolhas e formação, direta ou indiretamente.

A todos os envolvidos, muito obrigado.

RESUMO

Nos últimos anos o Brasil passou por um grande avanço na bancarização da população, partindo de 60% dos adultos em 2005, para 89,9% em 2019. Embora o setor financeiro venha dando grandes saltos com a tecnologia, o planejamento financeiro é um problema de praticamente toda a família brasileira. Mesmo que já existam soluções tecnológicas que possam auxiliar a população em seu controle financeiro ainda existe uma baixa aderência, alcançando 45% da população brasileira. Um dos motivos para a falta de aderência aos aplicativos de gestão financeira estaria ligado à baixa qualidade no registro dos dados. Embora os recursos que utilizam a voz venham se tornando mais populares, em vista da agilidade que proporcionam, os aplicativos de gestão financeira fazem pouco uso deste recurso e mesmo os que possuem, são limitados, sendo a maior parte deles para conversão de voz em texto para algum fim descritivo e não um comando que realize uma ação. Este trabalho desenvolveu, através da criação de artefatos, utilizando DSR (*Design Science Research*), um aplicativo que além de realizar os devidos registros financeiros, pode realizá-los por comandos de voz. Com uma taxa de assertividade de aproximadamente 70%, o usuário pode realizar seus registros financeiros usando a fala, necessitando apenas revisar o registro sugerido. Este recurso, pouco explorado atualmente, pode de fato trazer a agilidade e praticidade aos aplicativos de gestão financeira para que eles gerem maior atratividade e aderência a sua utilização.

Palavras-chave: Finanças pessoais, Aplicativo, Comandos de Voz, *Machine Lerning*, *Information extraction*.

ABSTRACT

In recent years, Brazil has experienced great progress in the use of banking services by the population, from 60% of adults in 2005 to 89.9% in 2019. Although the financial sector has been making great leaps with technology, financial planning is a problem practically the entire Brazilian family. Even if there are already technological solutions that can help the population with their financial control, there is still a low adherence, reaching 45% of the Brazilian population. One of the reasons for the lack of adherence to financial management applications would be linked to the low quality of data recording. Although resources that use voice are becoming more popular, in view of the agility they provide, financial management applications make little use of this resource and even those that have it are limited, most of them for voice-to-text conversion. For some descriptive purpose and not a command that performs an action. This work developed, by creating artifacts, using DSR (Design Science Research), an application that, in addition to carrying out the proper financial records, can carry them out by voice commands. With an assertiveness rate of approximately 70%, the user can carry out their financial records using speech, only needing to review the suggested record. This resource, currently little explored, can in fact bring agility and practicality to financial management applications so that they generate greater attractiveness and adherence to their use.

Keywords: Personal finances, Application, Voice Commands, Machine Lerning, Information extraction.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE SIGLAS

AED	*Attention-based Encoder-Decoder*
API	Interface de Programação de Aplicações
AR	*Adversarial Regularization*
Bacen	Banco Central do Brasil
CTC	*Connectionist Temporal Classification*
CNC	Confederação Nacional do Comércio
DAT	Domain Adversarial Training
DNN	Deep Neural Network
DSR	*Design Science Research*
FGSM	*Fast Gradient Sign Method*
GAN	*Generative Adversarial Network*
HMM	*Hidden Markov Model*
LAS	*Listen, Attend and Spell*
LM	Modelo de Linguagem
MFCC	Coeficientes Cepstrais de Frequência Mel
ML	*Machine Learning*
OS	*Operating System*
PCM	*Pulse Code Modulation*
RNN	Redes Neurais Recorrentes
RS	Revisão Sistemática
SPI	Sistema de Pagamento Instantâneo
TCC	Trabalho de Conclusão de Curso
URA	Unidades de Resposta Audível
ZCPA	Cruzamento por Zero com Amplitude Pico

SUMÁRIO

1. INTRODUÇÃO

Nos últimos anos o Brasil passou por um grande avanço na bancarização da população, partindo de 60% dos adultos em 2005, para 89,9% em 2019 (Caparelli; Spinola, 2020 apud CAVALCANTI; FILHO, 2021). Aproximando o Brasil “de países economicamente avançados como os Estados Unidos, com uma população bancarizada de 93,5%” (Vargas; Santos, 2020 apud CAVALCANTI; FILHO, 2021).

O cenário econômico vem crescendo com diversas transformações, fugindo das barreiras físicas, migrando para o meio digital e se tornando globalizado. As pessoas têm interagido cada vez mais com os meios digitais e utilizando ferramentas para facilitar seu dia a dia, estando estas ao alcance da maioria da população (VANDERLEY et al, 2020).

A evolução tecnológica, especialmente a miniaturização, vem auxiliando muito para esta transformação, principalmente a dos aparelhos móveis que antes eram usados para ligações e troca de mensagens de texto, permitindo agora a transferência de dados; muitos aplicativos já vêm interligados ao sistema operacional dos aparelhos comprados, ou mesmo podem ser adquiridos para execução de tarefas específicas. (CAVALCANTI; FILHO, 2021). Segundo APP ANNIE (2021 apud CAVALCANTI; FILHO, 2021), em 2020 foram adquiridos 218 bilhões de aplicativos no mundo, sendo 10 bilhões no Brasil.

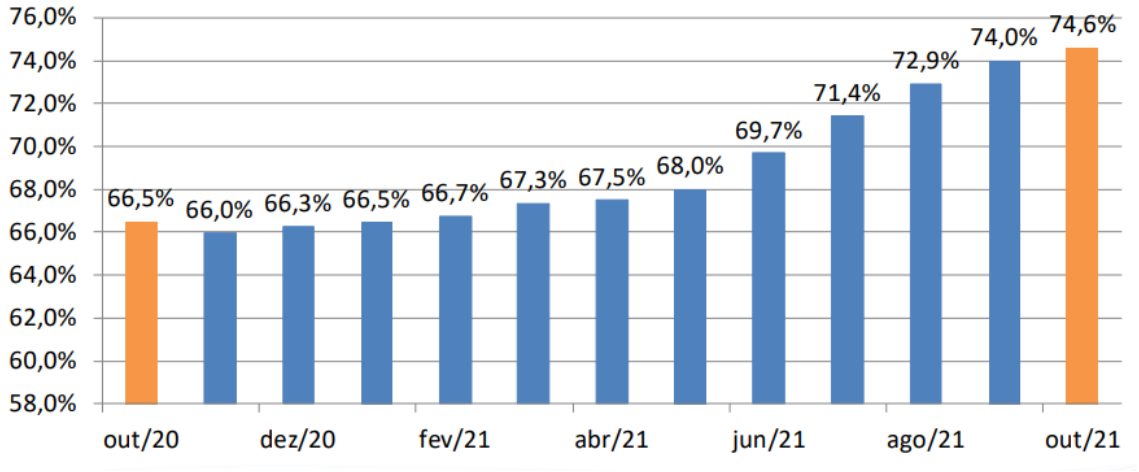
Embora o setor financeiro venha dando grandes saltos com a tecnologia, o planejamento financeiro é um problema de praticamente toda a família brasileira. Segundo Cerbasi (2014 apud COSTA et al, 2021), as pessoas se dedicam mais para as exigências profissionais do que as da vida pessoal e objetivos próprios. Outro fator é a falta de controle emocional frente a aquisições, cedendo frente a sedução de ofertas imperdíveis, algumas vezes se a pessoa fosse mais criteriosa perceberia a falta de necessidade da compra realizada.

Um exemplo dessa situação são os cartões de crédito, muito utilizado pelos jovens, estes trazem grande comodidade e uma sensação de despreocupação, embora seja necessário controlar as emoções e impedir comprar por impulso. Caso não haja o devido controle, no final do mês, o jovem terá contas que não conseguirá pagar (VANDERLEY et al, 2020).

Segundo uma pesquisa realizada pela Confederação Nacional do Comércio (CNC) em outubro de 2021, 74,6% dos brasileiros estão endividados, sendo que

25,6% possuem dívidas ou contas em atraso e pelo menos 10,1% não terão condições de pagar. O gráfico da Figura 1 mostra um aumento do índice do final de 2020 até outubro de 2021.

Figura 1 – Percentual de famílias endividadas (% do total)
(cartão de crédito, cheque especial, cheque pré-datado, crédito consignado, crédito pessoal, carnê de loja, prestação de carro e prestação de casa)



Fonte: CNC – Endividamento e Inadimplência do Consumidor, Out/2021

Visando atender as soluções e problemas de mercado, empresas estão investindo em aplicativos de finanças. Segundo Google (2021 apud CAVALCANTI; FILHO, 2021) estes aplicativos podem ser divididos em três grandes grupos: aplicativos disponibilizados por instituições financeiras, de pagamento e *cripto exchange*, destinados a realização de transações de recursos, pagamentos e produtos bancários; aplicativos para fornecimento e consolidação de informações sobre o mercado financeiro; aplicativos de gestão financeira para a consolidação de gastos e finanças.

Segundo Cavalcanti e Filho (2021):

> Os primeiros aplicativos para organização ou gestão financeira pessoal surgiram em 2012 (SILVA et al., 2018, p.2). Desde então, foram mais de 17 milhões de aplicativos adquiridos, destacando-se com o maior número de aquisições na loja de aplicativos da Google: GUIABOLSO, com mais de 10 milhões de downloads, MOBILLS, com mais de 5 milhões de downloads; MINHAS ECONOMIAS, com mais de 1 milhão de downloads; e ORGANIZZE, com mais de 1 milhão de downloads (GOOGLE, 2021). Esses números demonstram que ainda existe uma grande parcela da população que não utiliza o aparelho celular como uma ferramenta de controle financeiro.

Embora já existam soluções que possam auxiliar a população em seu controle financeiro ainda existe uma baixa aderência, “alcançando 45% da população brasileira (CNDL, 2018 apud CAVALCANTI; FILHO, 2021)”. Um dos motivos para a falta de

aderência aos aplicativos de gestão financeira estaria ligado a baixa qualidade no registro dos dados. Segundo Google (2021 apud CAVALCANTI; FILHO, 2021), a falta de integração destes aplicativos com contas bancárias e de pagamento, resulta em um cadastro manual pelo utilizador, o que não diferencia de um registro manual em planilha ou caderno de contas.

Uma das principais funcionalidades deste trabalho de conclusão de curso, está ligada ao reconhecimento de instruções por voz, assim os registros serão feitos de forma ágil. Segundo Redação (2019), o Google vem utilizando técnicas de *machine learning* desde 2014 com a assistente virtual, e esta é uma das técnicas que deixarão as soluções mais ágeis, sofisticadas, seguras e até mesmo capazes de funcionar *offline*.

Segundo Souza (2019), os comandos por voz tornam a vida mais fácil, trazendo praticidade no dia a dia, o que vai ao encontro do que é relatado por Google (2014 apud Souza 2019), onde “55% dos adolescentes usam pesquisas por voz pelo menos uma vez por dia e 89% concordam que as assistentes virtuais são o futuro”.

Grandes empresas estão investindo cada vez mais em inteligência artificial, pois veem novas perspectivas de serviços, assim como melhorias tanto internas como externas, seja na comunicação, produtividade, gerência etc. Muitas destas pensando justamente em “voice first” (Souza, 2019).

Segundo Cuadros (2007), “a comunicação vocal entre pessoas e máquinas engloba a síntese de texto para voz, reconhecimento automático de voz (conversão voz-texto), o reconhecimento de locutor e a codificação da voz”. Para que estes reconhecimentos ocorram são utilizados modelos não lineares, tais como MFCC (Coeficientes Cepstrais de Frequência Mel) e ZCPA (Cruzamento por Zero com Amplitude Pico), que procuram se aproximar do funcionamento do ouvido humano.

Este trabalho está dividido em 6 capítulos. Além da introdução, o capítulo 2 exibe uma pesquisa sobre as principais características pretendidas no aplicativo deste trabalho. O Capítulo 3 apresenta uma revisão sistemática de entendimento sobre as técnicas atuais para o desenvolvimento de aplicações com *machine learning* e comandos de voz. Os Capítulos 4 e 5 apresentam o desenvolvimento de uma aplicação financeira, com recursos de voz. Por fim, no capítulo 6, são feitas as considerações finais sobre o estudo realizado.

Durante o desenvolvimento do presente trabalho ocorreu muita pesquisa e troca de técnicas para contemplar os objetivos desejados. Ao termino foi possível

contemplar o objetivo principal, realizando registros utilizando apenas áudio (assertividade de 70%). Também foram abordados alguns dos específicos como a conversão de voz em texto e disponibilização das soluções em micro serviços. Nem todos os específicos puderam ser abordados pois devido a complexidade o presente estudo chegou apenas ao ponto do desenvolvimento do APP Financeiro, os demais objetivos partem deste ponto, sendo a análise de dados gerados e desenvolvimento de novos recursos específicos para o APP.

2. REFERENCIAL TEÓRICO

Para o desenvolvimento do presente trabalho foi definida a seguinte estratégia: no referencial teórico serão abordadas as principais características que o aplicativo deve conter. Com o objetivo de auxiliar na tomada de decisão das técnicas e tecnologias a serem utilizadas, o próximo capítulo abordará a revisão sistemática.

2.1. O crescimento dos meios de pagamento

Os meios de pagamento no Brasil vêm sofrendo grandes transformações, “a volumetria de transações envolvendo instrumentos de pagamento/cartões de crédito e débito cresceu mais de 500% durante os últimos anos” (Abecs, 2018 apud SOUSA W, 2021).

A Figura 2 retirada da pesquisa “O brasileiro e sua relação com o dinheiro” disponibilizada pelo Banco Central do Brasil (Bacen) em 2018, demonstra os principais meios de pagamento usados pelos brasileiros:

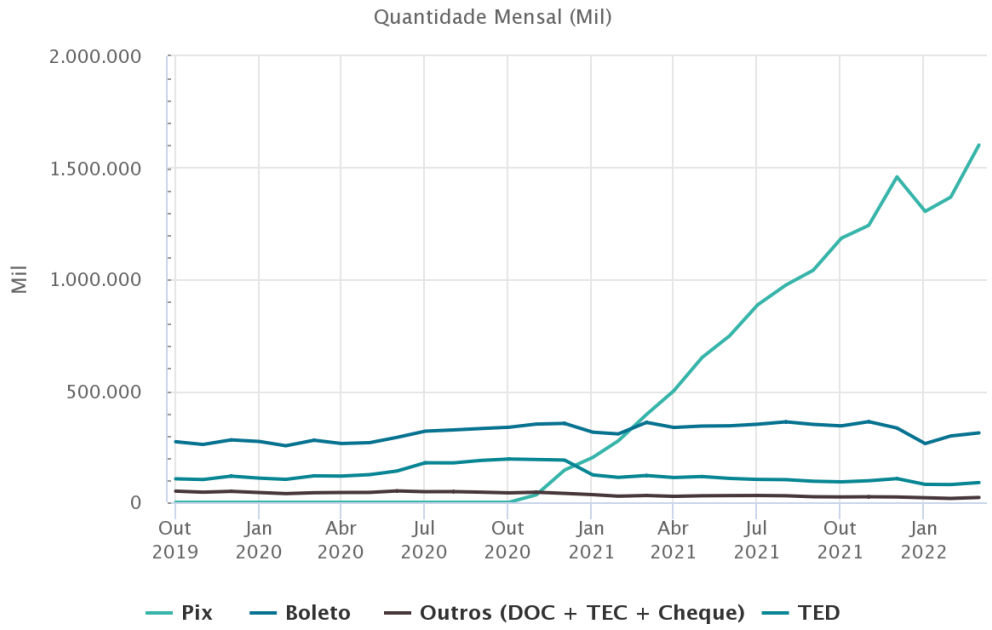
Figura 2 – Formas de pagamento adotadas para pagar as contas e/ou fazer compras

Fonte: O brasileiro e sua relação com o dinheiro (Bacen, 2018)

Já em novembro de 2020 o Bacen lançou o PIX, um novo meio de pagamento que funciona como Sistema de Pagamento Instantâneo (SPI). Este foi muito bem aceito pelo público e sua utilização vem crescendo substancialmente. Segundo Bacen (2020), através de dados estatísticos, o PIX em um ano já alcançou 1,2 milhões de transações, assim como está passando a ser um dos principais meios de pagamento,

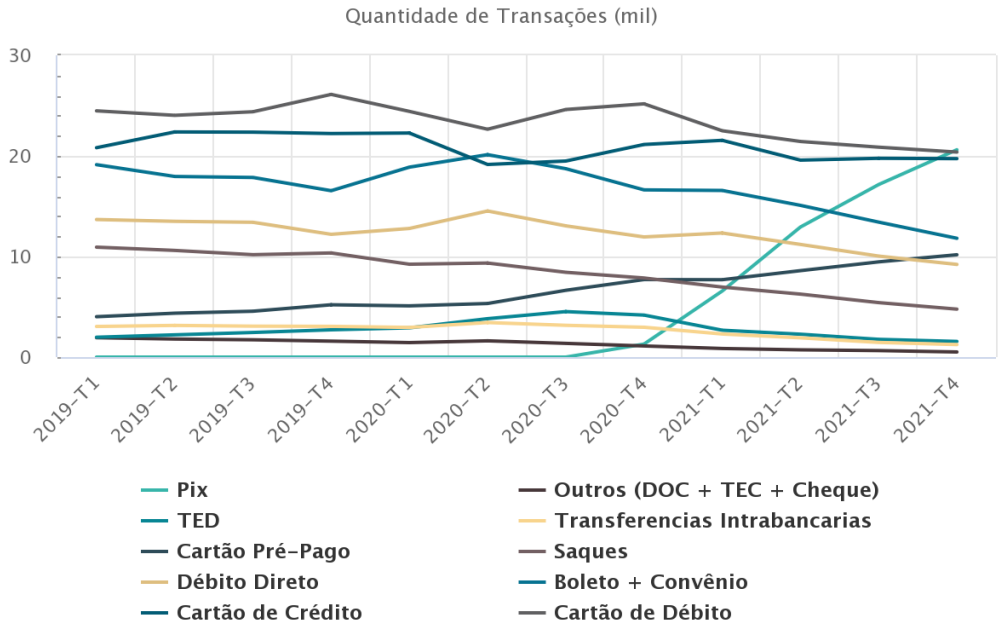
alcançando 20% da participação dentre os já existentes. Os gráficos das Figura 3 e Figura 4 demonstram estas informações:

Figura 3 – Dados Mensais - Meios de Pagamentos e Transferências



Fonte: Estatísticas de Meios de Pagamentos (Bacen, 2020)

Figura 4 – Dados Trimestrais - Participação Percentual Por Instrumento



Fonte: Estatísticas de Meios de Pagamentos (Bacen, 2020)

Ainda no gráfico da Figura 4 podemos perceber um grande número de meios de pagamento disponíveis, todos estes integram o planejamento financeiro pessoal.

2.2. O crescimento dos aplicativos de finanças pessoais

Segundo Coutinho (2014, p. 12, apud CAVALCANTI; FILHO, 2021), aplicativos podem ser entendidos como *softwares* produzidos para serem aplicados em um sistema operacional de um aparelho eletrônico móvel, como *tablets* e celulares *smartphones*. Atualmente a quantidade de aplicativos disponíveis em lojas virtuais é tão grande que possui uma economia própria, gerando empregos e renda diretamente ligados aos aplicativos. Em 2020 foram adquiridos 218 bilhões de aplicativos no mundo, sendo 10 bilhões no Brasil (APP ANNIE, 2021, apud CAVALCANTI; FILHO, 2021).

Existem três grandes grupos para classificar aplicativos relacionados a finanças: aplicativos de instituições financeiras, de pagamento ou manipulação de ativos, destinados a realizar transferência de recursos, pagamentos ou compra de produtos bancários; aplicativos de informação sobre o mercado financeiro; e aplicativos de consolidação de gastos e gestão financeira (Google, 2021, apud CAVALCANTI; FILHO, 2021).

Alguns aplicativos de gestão financeira foram identificados no Google Play, segundo informações disponibilizadas nas páginas de cada um dos aplicativos, com as seguintes datas de lançamentos: Minhas Economias, 2012; Meu Dinheiro, 2012; Minhas Finanças, 2013; Mobills, 2013; Guia Bolso, 2015; e Organizze, 2015. Podemos observar uma investida ao longo dos anos na criação destes meios de auxiliar as pessoas com a gestão pessoal.

O Quadro 1 mostra os principais recursos encontrados nos aplicativos do mercado atual.

Quadro 1 – Recursos encontrados nos aplicativos de finança no mercado atual

Recurso	**Guia Bolso**	**Meu dinheiro**	**Minhas Economias**	**Minhas Finanças**	**Mobills**	**Organizze**
Conta (valores por carteira)	x	x	x	x	x	x
Categorias (Farmácia, Mercado, etc)	x	x	x	x	x	x
Tags (Subcategorizador)		x		x	x	x
Receitas	x	x	x	x	x	x
Despesas	x	x	x	x	x	x
Transferências			x	x	x	
Cartão de Crédito	x	x	x	x	x	x
Integração com bancos	x		x	x	x	x
Relatórios	x	x				x
Exportar para Excel				x	x	

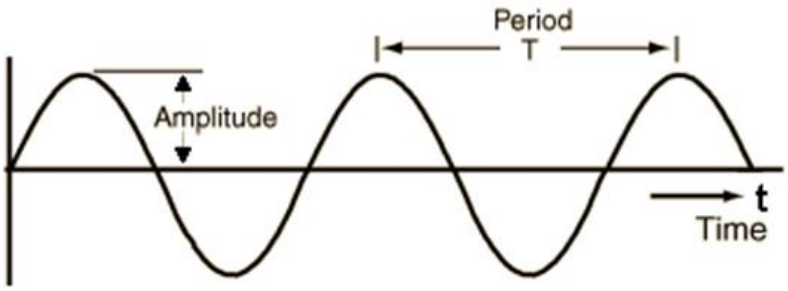
Recurso	Guia Bolso	Meu dinheiro	Minhas Economias	Minhas Finanças	Mobills	Organizze
Gráfico de Receita vs Despesas por categoria		x	x		x	
Repetidor			x			
Alerta / Lembrete		x	x	x	x	x
Planejador (% que deseja gastar em cada categoria vs % até o momento)					x	
Objetivos (Rotina para ajudar guardar dinheiro)	x			x	x	
Calendário (Sistema de paginação por mês)	x	x	x	x	x	x
Localização (Usa a localização para gravar onde foi registrado)					x	
Sistema de Fala						
Versão Web		x	x		x	x
Conversor de moedas				x		
Empréstimo	x					
Consulta CPF	x					
Gratuito	x	x	x	x	/	/
Lançamento	2015	2012	2012	2013	2013	2015

Fonte: Elaborado pelo autor

2.3. Digitalização de Voz

A voz ou também um sinal sonoro, é produzido pela variação na pressão do ar. A medição contínua deste sinal em relação ao tempo gera o que chamamos de sinal analógico. Na Figura 5 a altura mostra a intensidade do som, sendo conhecida como amplitude, já o tempo que uma onda leva para completar um ciclo, retomando seu início, é chamado de período e o número de ondas realizadas em um mesmo segundo (tempo) é conhecido como frequência ou *Hertz* (DOSHI, 2021).

Figura 5 – Sinal mostrando a amplitude e período sobre o tempo



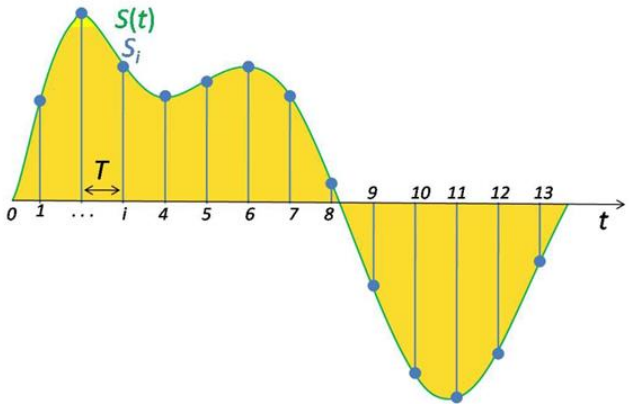
Fonte: Ketan Doshi (2021)

Segundo Cisco (2006), o primeiro processo que a digitalização de voz passa é a conversão de um sinal analógico para digital. Este, por sua vez, envolve três etapas principais: amostragem, quantização e compressão, esta última sendo optativa.

A etapa de amostragem é a gravação ou captura de um período de voz e a quantização é o processo de atribuir valores numéricos à amplitude de cada uma das amostras sendo codificada em valores em binários. O processo de compressão, etapa optativa, ocorre ao final da quantização, onde os valores passam por um algoritmo/técnica que diminui a quantidade de caracteres envolvidos para sua representação, por exemplo, converter binários para hexadecimal, onde “11000100” (binário) pode ser representado apenas com “c4” (hexa).

A modulação por código de pulso (do inglês *Pulse Code Modulation* - PCM), é um dos meios de conversão de sinal analógico para digital e se baseia em coletar 8000 amostras por segundo, com codificação binária de 8 bits. Na Figura 6 temos um exemplo mais simples de medição ao longo do tempo.

Figura 6 – Valoração de amostra ao longo do tempo



Fonte: Ketan Doshi (2021)

A onda desenhada em amarelo trata de uma captura de som no formato analógico. Esta foi dividida em intervalos de tempo “T” regulares (exemplo, a diferença entre cada um dos pontos no eixo X sempre será 1 segundo). Em relação aos tempos “t” (ponto do eixo X) é definido o maior valor da amplitude (conforme os pontos em azul) no momento, assim S(t) é igual a amplitude máxima no momento “t”.

2.3.1. Espectrograma do som

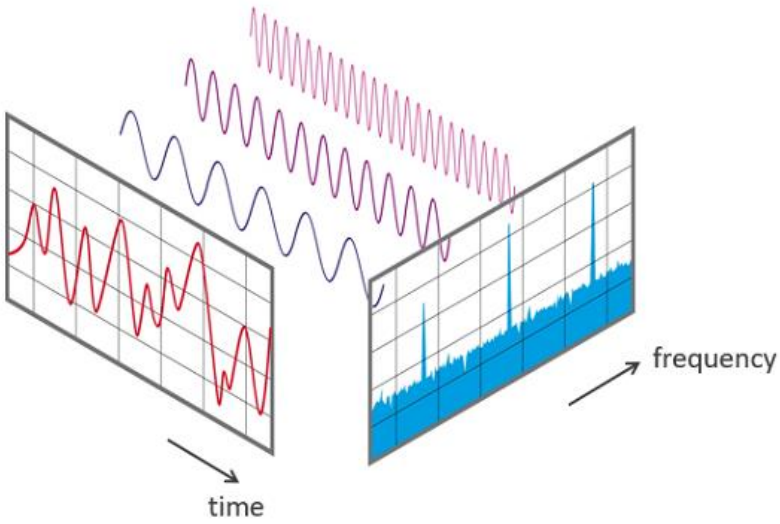
Diferente de melodias e alarmes, que seguem períodos regulares, em sua maioria o som presente no ambiente é na verdade uma composição de diferentes

frequências, ou seja, o sinal capturado também será uma combinação de muitas frequências distintas.

Um espectrograma mostra todas as frequências encontradas em um sinal, assim como, a força de cada uma delas. À medida que o tempo passa o sinal também captura sons diferentes, o que significa que as frequências e intensidades também variam. Segundo Doshi (2021), “Um espectrograma de um sinal traça seu espectro ao longo do tempo e é como uma 'fotografia' do sinal. Ele plota o tempo no eixo x e a frequência no eixo y”.

Na Figura 7 são exibidas diversas informações, na parte esquerda (em vermelho) observa-se a captura de um sinal em um ambiente, este trata de uma composição, pois não possui períodos regulares. Na parte interna (variando em tons de rosa) observa-se sinais regulares que estão sendo emitidos no ambiente, estes possuem frequências distintas. Na parte direita (em azul), há as frequências encontradas no ambiente, assim como, sua amplitude (força).

Figura 7 – Captura de sinal, composição e frequência



Fonte: Ketan Doshi (2021)

2.3.2. A tecnologia da voz

A tecnologia da voz vem sendo usada desde a criação do telefone por Alexander Grahan Bell, e está sofrendo melhorias, passando dos pulsos elétricos nas vias telefônicas, para radiofrequência com os telefones sem fio, até ondas eletromagnéticas com os smartphones (TECNOLOGIA POSITIVO, 2019).

O principal objetivo desta tecnologia sempre foi de facilitar a comunicação humana, atualmente passando a facilitar até mesmo a comunicação humano-máquina.

Com o surgimento da indústria 4.0 esta tecnologia vem ganhando cada vez mais aderência, alguns exemplos de sua aplicação são: as Unidades de Resposta Audível (URA), muito utilizadas na rede de *call center*, estas atendem automaticamente a chamada do cliente e direcionam para a fila ou área desejada; outro impacto bastante significativo, foram as buscas por voz, que segundo Google (2014 apud Souza 2019): “cerca de 20% de todas as consultas são pesquisas por voz em dispositivos móveis” e ainda segundo ele “projeções mostram aumento para 50% [...] até 2020”, mas existem diversos outros recursos já disponíveis como, pagamentos por voz, liberação de acesso, controle de objetos, reconhecimento de voz, entre outros.

2.3.3. O futuro e os recursos de voz

A utilização de assistentes virtuais, URA, chatbots, dentre outros recursos trazem melhorias para o atendimento aos clientes, assim como, minimizam os erros internos dentro das empresas, trazendo impactos aos custos operacionais de quem os utiliza.

Segundo uma pesquisa realizada pela Jupiter Research (Moar J, Escherich M, 2020), apenas o uso de chatbots em bancos economizará até US$ 5,7 bilhões em todo o mundo, somente em 2024. Segundo o mesmo estudo, até 2024 “os consumidores irão interagir com assistentes de voz em 8,4 bilhões de dispositivos”.

Segundo Oh S et al (2021), o número de dispositivos de infroentretenimento em veículos vem aumentando, estes com o objetivo de aumentar a satisfação e fornecer novas funções de controle ao veículo. Alguns destes, porém, podem fornecer distrações ao condutor, como a configuração de um sistema de GPS ou entretenimento de áudio. Para tentar mitigar estas interações manuais, se utilizam sensores de voz que possam compreender ações que o veículo possa tomar, sem a necessidade de interação direta pelo usuário.

O surto de Coronavírus (COVID-19) foi anunciado como uma pandemia global pela Organização Mundial da Saúde em março de 2020 (Shimon C et al, 2021), com esse cenário diversas tentativas de auxiliar nos diagnósticos foram realizadas. Segundo Han et al (2020 apud Shimon C et al, 2021), como o COVID-19 é uma

doença respiratória, os padrões respiratórios anormais dos pacientes intuitivamente podem ser um indicador potencial para o diagnóstico da qualidade do sono, ansiedade, fadiga. O estudo realizado por Shimon C et al, demonstrou a viabilidade e a eficácia da análise de COVID-19 baseada em áudio texto.

Estes estudos mostram um grande crescimento na utilização desta tecnologia, que tende a ficar cada vez mais robusta e fazer parte no dia a dia das pessoas.

2.4. *Machine learning* e Inteligência Artificial

Machine Learning (ML) é o termo em inglês para uma tecnologia conhecida no Brasil como Aprendizado de Máquina fazendo parte da Inteligência Artificial (IA), que é todo um campo do conhecimento, semelhante à biologia ou à química. Ela surgiu com o objetivo de prever resultados com dados históricos armazenados, portanto é importante possuir a maior variedade de amostras nos históricos para assim encontrar mais facilmente padrões relevantes e prever os resultados discretos ou contínuos.

Para ensinar a máquina são necessários três componentes essenciais: dados, recursos e algoritmos. Quanto mais dados, melhor será o resultado. Por exemplo, se você quer descobrir as preferências dos usuários, analise suas atividades nas redes sociais, obtendo a maior variedade de dados possível. A obtenção de um conjunto de dados pode ser manual ou automática, sendo a segunda mais rápida e barata ao contrário da primeira. Já os recursos são fatores que a máquina precisa observar, continuando no exemplo das preferências dos usuários, os recursos também conhecidos como parâmetros ou variáveis dos usuários, neste caso seriam: sexo, idade, profissão, poder de compra, renda, entre outros. No caso do algoritmo para analisar tais dados, não adianta dispor do melhor hardware ou software, se seus dados forem ruins. É possível observar que cada etapa depende da outra para a obtenção do melhor resultado no final.

A utilização de *machine learning* vem ganhando cada vez mais aderência, alguns exemplos práticos são: reconhecimento de imagens – sendo mais comumente usado para classificar objetos e reconhecer faces; reconhecimento de voz – onde o software tem a capacidade de reconhecer palavras e converter em texto, usado em atendimentos de URAs; indicações de filmes, música e conteúdo em geral – utilizado pelos serviços de *streaming* (como Netflix e Youtube), formando as recomendações (Oliveira F, 2019).

2.4.1. Objetivo de *Machine Learning*

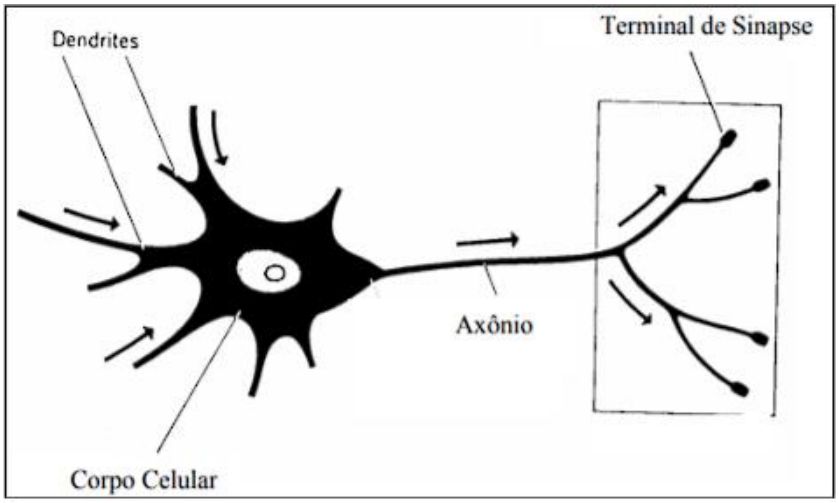
O desenvolvimento de ML se deu com o objetivo de tornar o computador capaz de aprender com a própria experiência passada. Esta área do conhecimento desenvolveu algoritmos capazes de obter conclusões a partir de um conjunto de exemplos e com estas conclusões, deduzir uma hipótese ou função capaz de resolver um problema baseado nos dados que demonstram iminência do problema. (Katti Faceli et al, 2011 apud Augusto C et al 2018).

2.4.2. Modelo de Neurônio Computacional

As Redes Neurais Artificiais (RNAs) são modelos criados com base nas redes de neurônios biológicos, como no cérebro humano. A Figura 8 demonstra um neurônio biológico, que possui as seguintes características:

- Dendrites: faz a coleta dos impulsos oriundos de outros neurônios;
- Corpo Celular: responsável por processar os sinais recebidos pelas dendrites;
- Axônio: encarregado pela multiplicação dos sinais;
- Sinapse: conexão do axônio de um neurônio com a dendrite de outro.

Figura 8 - Neurônio Biológico



Fonte: Bianchini 2004 apud Augusto C et al 2018

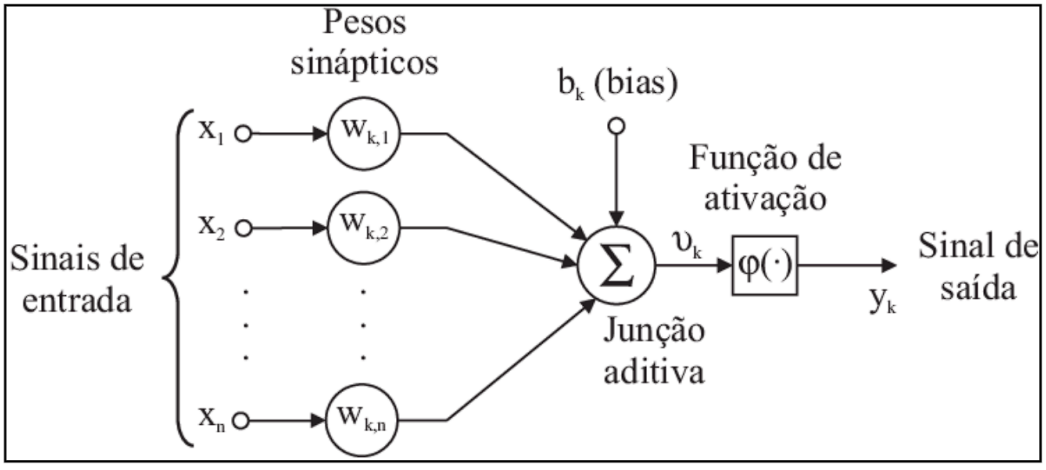
Inspirados pela observação das interações dos neurônios biológicos, os cientistas de dados desenvolveram o Modelo de Neurônio Artificial, que pode ser observado na Figura 9, este neurônio artificial possui as seguintes características:

- Conjunto de sinapses: como visto no modelo biológico, a sinapse é a responsável por ligar um neurônio a outro. Neste caso, o conjunto de sinapses

é composto pelos sinais de entrada e pesos sinápticos, de modo que, cada sinal de entrada será multiplicado pelo peso sináptico associado a ele. Assim, o neurônio irá tratar de maneira diferente cada impulso recebido, atribuindo-lhe um valor;

- Um somador: é responsável por somar os sinais de entrada, ponderados pelo respectivo valor de peso;
- Bias: é responsável por acrescentar algum viés ao aprendizado. É ideal que o valor do bias seja próximo de zero, assim o modelo não possui viés, sendo mais genérico, entretanto o modelo pode acabar tendo uma grande variância, tornando-se impreciso. O bias ajuda a encontrar um equilíbrio para a generalização desejada.
- Função de ativação: é utilizada para restringir a amplitude do sinal de saída do neurônio.
- Sinal de Saída: é o resultado da ponderação dos pesos juntamente com o bias, seu valor é uma estimativa de precisão em relação a generalização dos dados. O modelo aprende uma função de generalização com o treinamento e a saída é uma estimativa de proximidade da entrada atual em relação a generalização aprendida.

Figura 9 - Neurônio Artificial Computacional



Fonte: Bianchini 2004 apud Augusto C et al 2018

O neurônio artificial é o pilar para as RNAs, que possuem diversas camadas. O sistema ao fazer o processamento da informação, altera os valores dos pesos dos neurônios presente na rede, adaptando sua própria estrutura a fim de chegar o mais próximo possível do resultado desejado (JOST, 2015 apud Augusto C et al 2018).

2.4.3. Aprendizado supervisionado e não supervisionado

Segundo Augusto C et al (2018), as técnicas de ML utilizam o princípio de inferência conhecido como indução, na qual o computador consegue obter conclusões genéricas a partir de um conjunto de exemplos, sendo o aprendizado indutivo obtido de forma supervisionada ou não-supervisionada.

No aprendizado supervisionado, um conjunto de dados sobre um determinado objetivo é previamente desenvolvido, este conjunto possui exemplos de entradas e saídas esperadas. “O algoritmo de ML adquire a representação do conhecimento com base nesses exemplos, a fim de que as representações geradas sejam capazes de produzir saídas corretas para novas entradas, não apresentadas previamente” (Augusto C et al, 2018).

O aprendizado não supervisionado é mais utilizado quando o objetivo é identificar padrões ou tendências, onde não existe um desenvolvimento prévio ou valores de referência. Este algoritmo aprende a representar ou agrupar seguindo medidas de similaridade (Augusto C et al, 2018).

2.4.4. Deep Learning

Deep Learning (DL) ou Aprendizado Profundo, é uma área de pesquisa extremamente ativa, sendo utilizada para reconhecimento de voz, processamento de imagem, mineração de dados, classificação de doenças, entre outras (COPELAND, 2016 apud Augusto C et al 2018).

Segundo Arnold et al (2011 apud Augusto C et al 2018), “O *Deep Learning* surgiu como o paradigma que trata a dificuldade de arquiteturas frequentemente utilizadas, [...] que possuem alta dimensão de dados”. Este algoritmo de múltiplas camadas, obteve resultados positivos na obtenção de características em dados não rotulados, na qual os dados das características de alto nível são fornecidas pela composição das de baixo nível, por exemplo, ao processar uma imagem com DL, a camada 1 fica responsável por agrupar *pixels* que possam formar uma reta, a camada 2 é responsável por buscar padrões geométricos com estas retas e a camada 3 é responsável por unir estes padrões geométricos, identificando um objeto. Neste exemplo a camada 3 é superior e dependente da camada 2, da mesma forma que a camada 2 é superior e dependente da camada 1.

Segundo Ponti e Costa (2017, apud Augusto C et al, 2018):

> A diferença entre DL e ML está na função $f(x)$, onde técnicas que não utilizam DL são frequentemente chamadas de “superficiais” ou “rasas” (*shallow*), pois buscam uma única função a partir de um grupo de parâmetros, gerando um resultado desejado. Porém, para o DL existem técnicas que aprendem a função a partir da composição de funções: $f(x)=fL(\cdots f2(f1(x1))\cdots))$, onde, para cada função $fl(.)$ o índice l refere-se a uma “camada”, o $x1$ tem como entrada um vetor de dados e como saída o vetor $xl+1$. As funções se utilizam de parâmetros para transformar dados de entrada. [...] como uma matriz Wl, referente a cada função fl, $fL(\cdots f2(f1(x1,W1);W2)\cdots),WL)$ onde $x1$ retrata os dados de entrada, cada função faz o uso do próprio conjunto $W1$ de parâmetros e sua saída será passada para a próxima função.

Pode-se concluir que a diferença entre DL e ML é que, enquanto o ML busca criar uma função com base em uma expressão algébrica simples, apenas alterando os pesos, o DL gera funções dependentes em que os resultados das camadas inferiores servem como informação para as camadas de nível superior.

2.4.5. Processamento de Linguagem Natural

Segundo Rodrigues J. (2017), “o processamento de Linguagem Natural (PLN) é a subárea da Inteligência Artificial (IA) que estuda a capacidade e as limitações de uma máquina em entender a linguagem dos seres humanos”. O objetivo do PLN é criar meios, algoritmos, para que os computadores possam “entender” o que os humanos falam ou escrevem.

Diferente de informações estruturadas, que seguem um determinado padrão, tornando mais simples criar meios para extrair dados, a fala ou escrita humana não é determinística, possuindo alto grau de complexidade, onde a mesma frase pode ser escrita de modos diferentes ou possuir um significado diferente dependendo do contexto em que está inserida. O “entender” um texto, significa “reconhecer o contexto, fazer análise sintática, semântica, léxica e morfológica, criar resumos, extrair informação, interpretar os sentidos, analisar sentimentos e até aprender conceitos” (Rodrigues J, 2017). Nos últimos anos o PLN esteve presente em soluções como classificação textual, análise de sentimentos, tradução automática, sumarização de textos (Soares De Paiva E, Pereira F, 2022), assistentes virtuais, buscadores, Unidades de Resposta Audível (URAs), entre outras.

O primeiro modelo a obter resultado satisfatório na análise de PLN, foi o desenvolvido por Mikolov et al em 2013, ao construir uma rede neural de duas camadas, ficando conhecida como Word2Vec. Este modelo trabalha com a ideia de palavras centrais, tendo duas variações “Skip-gram” e “Continuous Bag of Words (CBOW)”. A primeira funciona gerando uma lista de possíveis palavras subsequentes,

dada uma palavra na entrada, a segunda gera a possível palavra, dada uma lista de palavras. (Paiva S E, Pereira S F, 2022).

2.5. Considerações finais do referencial teórico

Podemos verificar, no incio deste capitulo, que o desenvolvimento de um aplicativo financeiro com recursos de voz, tende a ser uma inovação visto seu potencial ainda inexplorado. Os recursos de voz têm crescido nos últimos anos, estando cada vez mais integrados às soluções existentes, facilitando o dia a dia das pessoas.

Entretanto, o desenvolvimento deste aplicativo não é uma tarefa simples. Embora já existam diversos aplicativos financeiros no mercado, o principal complicador está justamente no recurso de voz que necessita de diversas tecnologias complexas, como a manipulação da voz e aprendizado de máquina.

3. REVISÃO SISTEMÁTICA

Com o intuito de obter um guia a fim de direcionar o desenvolvimento da aplicação futura, foram realizados alguns passos de uma revisão sistemática. Tais passos têm o objetivo de obter *insights* sobre o tema deste TCC e descobrir aplicações similares a ele, assim como evitar diretrizes que não obtiveram resultados favoráveis em outros trabalhos.

Com base nos resultados desta revisão serão selecionadas técnicas, linguagens de programação, ferramentas e/ou frameworks que possam ser utilizados para o desenvolvimento do aplicativo.

3.1. Metodologia da revisão sistemática

Neste estudo será abordado o conceito da revisão sistemática (RS) proposto por (ROEVER L, 2017) na qual segundo ele: “A revisão sistemática consiste em um processo de pesquisar, selecionar, avaliar, sintetizar e relatar as evidências clínicas sobre uma determinada pergunta e/ou tópico”. Neste caso adaptando para obtenção das melhores diretrizes de desenvolvimento da aplicação.

3.2. Estrutura da revisão sistemática

Seguindo a estrutura proposta por (ROEVER L, 2017) existem as seguintes definições:

3.2.1. Título

Revisão sistemática sobre aprendizado de máquina voltado a voz aplicado a recursos de finanças.

3.2.2. Resumo

A presente revisão sistemática foi realizada sobre a base de dados “ACM Digital Library” com foco em artigos que possuam relação com conversão de voz em texto, assim como inteligência artificial ligada a finanças. Com o objetivo de obter informações e técnicas que possam ser utilizadas no desenvolvimento de um aplicativo de finanças pessoais. Os resultados obtidos foram extremamente satisfatórios para entendimento das técnicas de conversão de voz em texto, assim como alternativas para o desenvolvimento do aplicativo proposto. Pode-se concluir

que o processo de utilizar uma revisão sistemática como base de conhecimento para iniciar uma aplicação é uma ótima opção.

3.2.3. Introdução

O desenvolvimento desta revisão sistemática tem como foco a obtenção das diretrizes para o desenvolvimento de uma aplicação de finanças com recursos de voz.

3.2.3.1. Objetivos

Obter diretrizes para o desenvolvimento de uma aplicação de finanças com recursos de voz.

3.2.4. Métodos

O modelo proposto por (ROEVER L, 2017), possui uma definição especifica para auxiliar no processo extração dos dados, que são:

3.2.4.1. Protocolo e registro

Não aplicável.

3.2.4.2. Critérios de elegibilidade

- Os estudos devem estar entre o período de 2017 e 2022;
- Ser um artigo publicado;
- Possuir relação com aprendizado de máquina para finanças e/ou voz;
- Apresentar técnica/framework que converta voz em texto;
- Apresentar quais técnicas e algoritmos foram utilizados com aprendizado de máquina;
- Nos estudos voltados a finanças: quais os tipos de finanças estão sendo abordadas;

3.2.4.3. Fontes de informação

Os artigos analisados serão obtidos da base de dados “ACM Digital Library” (https://dl.acm.org/search/advanced), que possui como base em revistas científicas (journals), anais da conferência, revistas técnicas, boletins informativos e livros. Sendo ainda um banco de dados bibliográfico focado exclusivamente no campo da computação.

3.2.4.4. Busca

Para a busca dos artigos foi criada a seguinte “*string* de busca”:

("speech to text" OR "text to speech") AND ("machine learning" OR "artificial intelligence") AND ("finance*" OR "bills" OR "personal expense" OR "personal spend*" OR "personal expenditure*")

3.2.4.5. Seleção dos estudos

Será a aplicação da *string* de busca na fonte de dados informada: Subsequente ao resultado obtido serão empregados os critérios de elegibilidade sobre o título, palavras-chave e resumo dos artigos encontrados, caracterizando a etapa 1. Com os artigos selecionados, na etapa 2 serão filtrados os artigos que obedecem aos critérios de elegibilidade a partir da leitura da Introdução e da Conclusão do artigo. Na etapa 3 será realizada a leitura completa deles.

3.2.4.6. Processo de coleta de dados

Com os artigos devidamente selecionados na etapa 3 será iniciado o processo de coleta das informações relacionadas aos critérios de busca. Demais informações que possam ser uteis para o desenvolvimento deste TCC serão incorporadas aos resultados finais da revisão sistemática.

3.2.4.7. Lista dos dados

Com objetivo de identificar as tecnologias empregadas, será criada uma tabela, a fim de pontuar os critérios relacionados aos artigos.

3.2.4.8. Risco de viés em cada estudo

Por se tratarem de estudos relacionados a tecnologia será considerado que as técnicas utilizadas foram as melhores esperadas para a solução e/ou aplicação da solução.

3.2.4.9. Medidas de sumarização

Não aplicável

3.2.4.10. Síntese dos resultados

Não aplicável

3.2.4.11. Risco de viés entre estudos

Não aplicável

3.2.4.12. Análises adicionais

Não aplicável

3.2.5. Resultados

A aplicação da *string* de busca na fonte de dados, conforme a cima informado, foi realizada no dia 16 de maio de 2022, para auxiliar no desenvolvimento deste estudo foi utilizada a ferramenta StArt (LaPES), focada justamente na realização de uma revisão sistemática. Seguindo as recomendações de (ROEVER L, 2017), os resultados foram:

3.2.5.1. Seleção dos estudos

Foram obtidos 33 artigos, conforme apresentado no Quadro 2:

Quadro 2 – Resultado da *string* de busca na fonte ACM

Título	**Ano**
Monitoring Misuse for Accountable 'Artificial Intelligence as a Service'	2020
Privacy-Preserving Machine Learning Based Data Analytics on Edge Devices	2018
Monitoring AI Services for Misuse	2021
SHIELD: A Framework for Efficient and Secure Machine Learning Classification in Constrained Environments	2018
Let Me Ask You This: How Can a Voice Assistant Elicit Explicit User Feedback?	2021
Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support	2022
Integrating Knowledge Into End-to-End Speech Recognition From External Text-Only Data	2021
Digital Agriculture for Small-Scale Producers: Challenges and Opportunities	2021
"You Know What to Do": Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks	2019
StoryDrawer: A Child - AI Collaborative Drawing System to Support Children's Creative Visual Storytelling	2022
Data Analytics Service Composition and Deployment on Edge Devices	2018
Adversarial Regularization for Attention Based End-to-End Robust Speech Recognition	2019
'Could You Describe the Reason for the Transfer?': A Reinforcement Learning Based Voice-Enabled Bot Protecting Customers from Financial Frauds	2021
Creating and Evaluating Chatbots as Eligibility Assistants for Clinical Trials: An Active Deep Learning Approach towards User-Centered Classification	2021
Commercialization of Multimodal Systems	2019
Unmet Needs and Opportunities for Mobile Translation AI	2020
The Handbook of Multimodal-Multisensor Interfaces: Language Processing, Software, Commercialization, and Emerging Directions	2019
DolphinAttack: Inaudible Voice Commands	2017
Global ICT Accessibility Methodologies for Persons with Disabilities and Initiatives in India	2017
The New Post-Pandemic Normal of College Traditions	2020
Tuning the ISA for Increased Heterogeneous Computation in MPSoCs	2020
Read Between the Lines: An Empirical Measurement of Sensitive Applications of Voice Personal Assistant Systems	2020
SoundSifter: Mitigating Overhearing of Continuous Listening Devices	2017
What's Wrong with Computational Notebooks? Pain Points, Needs, and Design Opportunities	2020

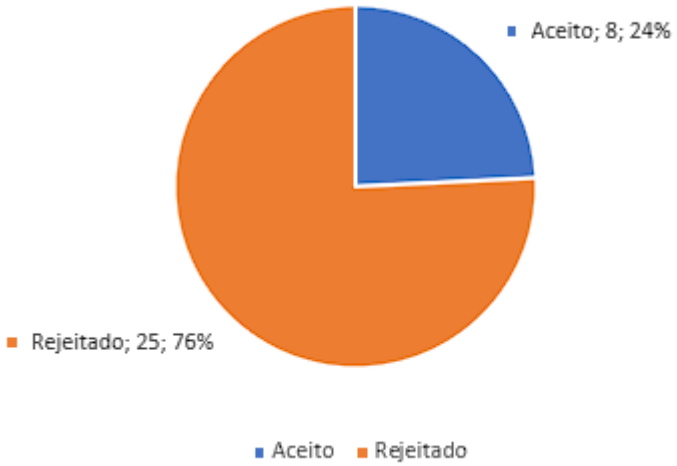
Smart Home Personal Assistants: A Security and Privacy Review	2020
WWW '19: Companion Proceedings of The 2019 World Wide Web Conference	2019
AIGuide: Augmented Reality Hand Guidance in a Visual Prosthetic	2022
The Design and Evaluation of a Mobile System for Rapid Diagnostic Test Interpretation	2021
CHI EA '21: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems	2021
Asymmetries in Online Job-Seeking: A Case Study of Muslim-American Women	2021
The Characteristics of Voice Search: Comparing Spoken with Typed-in Mobile Web Search Queries	2018
TVX '19: Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video	2019
CSCW '17: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing	2017

Fonte: Elaborado pelo autor

3.2.5.2. Etapa 1

Subsequentemente aplicou-se a filtragem seguindo os critérios de busca, sobre título, resumo e palavras-chave. Como resultado obteve-se uma redução de 76% dos resultados, conforme exibido na Figura 10.

Figura 10 – Classificação dos estudos na etapa 1.



Fonte: Elaborado pelo autor

Apesar de terem sido encontrados 33 resultados, alguns eram livros ou conferências, não sendo caracterizados como artigo para estudo. No Quadro 3 podemos observar os 8 estudos aceitos, representando apenas 24% do total.

Quadro 3 – Estudos selecionados após a etapa 1.

Título	**Ano**
Integrating Knowledge Into End-to-End Speech Recognition From External Text-Only Data	2021
StoryDrawer: A Child - AI Collaborative Drawing System to Support Children's Creative Visual Storytelling	2022
Adversarial Regularization for Attention Based End-to-End Robust Speech Recognition	2019

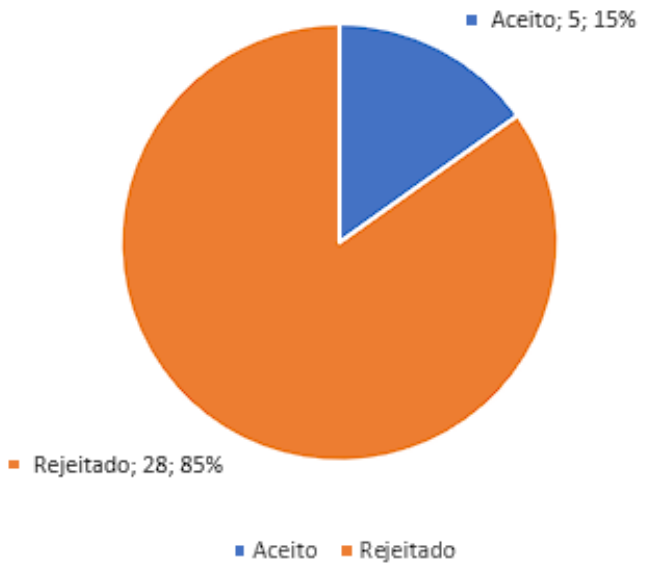
DolphinAttack: Inaudible Voice Commands	2017
Read Between the Lines: An Empirical Measurement of Sensitive Applications of Voice Personal Assistant Systems	2020
SoundSifter: Mitigating Overhearing of Continuous Listening Devices	2017
Smart Home Personal Assistants: A Security and Privacy Review	2020
The Characteristics of Voice Search: Comparing Spoken with Typed-in Mobile Web Search Queries	2018

Fonte: Elaborado pelo autor

3.2.5.3. Etapa 2

Sob os artigos selecionados na etapa 1, aplicaram-se os critérios de busca sobre introdução e conclusão. Nesta etapa, mais alguns artigos foram filtrados em vista que a metodologia não se adequava aos critérios. Desta forma, a redução chegou a 85%, conforme exibido na Figura 1.

Figura 11 – Classificação dos estudos na etapa 2.



Fonte: Elaborado pelo autor

Ao final deste processo restaram 5 artigos, que podem ser observados no Quadro 4, representando 15% dos 33 artigos iniciais.

Quadro 4 – Estudos selecionados após a etapa 2.

Título	**Ano**
Integrating Knowledge Into End-to-End Speech Recognition From External Text-Only Data	2021
StoryDrawer: A Child - AI Collaborative Drawing System to Support Children's Creative Visual Storytelling	2022
Adversarial Regularization for Attention Based End-to-End Robust Speech Recognition	2019
DolphinAttack: Inaudible Voice Commands	2017
SoundSifter: Mitigating Overhearing of Continuous Listening Devices	2017

Fonte: Elaborado pelo autor

3.2.5.4. Etapa 3

Na etapa 3 foi realizada a leitura completa dos artigos selecionados na etapa 2 e realizada a filtragem sobre o conteúdo apresentado, novamente seguindo os critérios de elegibilidade. Nesta etapa apenas um artigo foi desclassificado: *DolphinAttack: Inaudible Voice Commands* (Zhang G, Yan C, Ji X et al. 2017), pois apesar de ter correlação com reconhecimento da fala, ele é focado na análise de *hardware*, verificando meios de interferir na entrada de dados, observando a distorção das frequências para tentar embutir comandos nocivos no equipamento.

Apesar de todas as informações sobre como o som se propaga e como os equipamentos o capturam e convertem o áudio em informações digitais, em vista de que este trabalho de TCC não terá nenhum impacto sobre o *hardware* utilizado, nenhuma informação pode ser coletada.

Ao final das três etapas de filtragem, quatro artigos foram selecionados para a extração de dados e serão discutidos a seguir.

3.2.6. Resultados dos estudos individuais

A seguir serão apresentados os estudos selecionados e as principais características de cada um deles, levando em consideração a relevância para o desenvolvimento deste TCC, seguindo os critérios de elegibilidade.

3.2.6.1. Integrating Knowledge Into End-to-End Speech Recognition From External Text-Only Data (Bai Y, Yi J, Tao J et al, 2021)

Os autores apresentam técnicas de conversão de voz em texto. A técnica tradicional, o hibrido de “Hidden Markov Models (HMMs)” com *Deep Neural Network* (DNN), conhecido como DNN-HMM, realiza o treinamento do modelo acústico e linguístico separadamente. Segundo os autores, esta técnica possui três problemas: a construção de um modelo léxico para cada idioma é complexa, exigindo conhecimento especializado; o treinamento de n-gramáticas exige uma grande quantidade de textos; e como os modelos são treinados separadamente o erro que ocorrer em um deles é acumulado para o outro.

Outra técnica apresentada neste artigo é a A*ttention-based Encoder-Decoder* (AED) juntamente com Redes Neurais Recorrentes (RNN), que utiliza um modelo de acústica e linguística. Esta técnica não necessita de um pronunciador léxico e nem uma busca em grafo de palavras. O problema deste modelo está no treinamento, que

necessita de dezenas de gigabytes de informação e é realizado de forma paralela. Isto o torna muito mais caro do que o modelo tradicional que usa apenas texto.

Inspirados pelo aprendizado professor-aluno o que os autores propõem é o desenvolvimento de um método que possa realiza a passagem de conhecimento de um modelo para o outro, utilizando a representação do que foi aprendido nos textos e transferindo para o AED.

Para validar o modelo proposto, foi realizado um comparativo com LSTM (L*ong Short-Term Memory*) baseado em RNN e um Modelo de Linguagem (LM) baseado em transformação. Para realizar o experimento foram escolhidos dois *datasets* de fala chineses, AISHELL-1 e AISHELL-2, assim como um subconjunto do CLMAD, caracterizado como texto externo.

Ao decorrer do processo os autores realizaram alguns ajustes nas técnicas, com objetivo de obter resultados melhores, contudo o modelo proposto obteve melhor resultado em análises do texto externo, com acurácia de 65%. Em relação a outras análises os resultados foram inferiores ou iguais aos demais modelos.

Neste artigo foram extraídas as seguintes informações:

- Hidden Markov Models (HMMs) juntamente com DNN, é o método tradicional para reconhecimento de fala, mas necessitava da construção de um léxico para cada idioma desejada, escalando rapidamente.
- Recentemente modelos de AED têm se tornado mais populares em vista de não necessitar de léxicos e buscas estáticas.

3.2.6.2. StoryDrawer: A Child - AI Collaborative Drawing System to Support Children's Creative Visual Storytelling (Zhang C, Yao C, Wu J et al, 2022)

O trabalho utiliza o modelo YOLOv3 para identificar os desenhos feitos por crianças, usando como base de dados o serviço “Quick Draw”[1]. Este também foi utilização para gerar as sugestões nos comandos de voz. Diferentemente do trabalho anterior, o foco foi puramente no desenvolvimento do aplicativo e demonstração de seus resultados, no desenvolvimento da criatividade das crianças.

Para o desenvolvimento do aplicativo foram utilizados alguns serviços do google cloud, assim como uma biblioteca de python (NLTK) para extrair as informações do texto.

[1] *Quick Draw* é um serviço feito pelo Google que utiliza uma rede neural para tentar identificar o desenho que foi feito. Seu lançamento foi em 2016.

Neste artigo foram extraídas as seguintes informações:

- APIs fornecidas pelo Google, para conversão de voz em texto e tradução de linguagem. A tradução foi importante para tornar a aplicação genérica, assim como facilitar na utilização do modelo escolhido.
- Utilização do pacote python NLTK para extração de informações do texto.

3.2.6.3. Adversarial Regularization for Attention Based End-to-End Robust Speech Recognition (Sun S, Guo P, Xie L et al, 2019)

Da mesma forma que no artigo da seção 3.2.5.3.1, os autores informam que a técnica tradicional para reconhecimento de voz em texto é o hibrido DNN-HMM, e pesquisadores vem tentando aprimorar esta técnica utilizando de técnicas "*end-to-end*" (ponta-a-ponta) como C*onnectionist Temporal Classification* (CTC) e A*ttention based Model*. Ambos são classificados como modelo de *Listen, Attend and Spell* (LAS) e consideram o reconhecimento de fala como uma tarefa de sequência a sequência (*seq-to-seq*), que mapeia as sequências de recursos de fala para sequências de rótulos de texto.

Segundo os autores, embora esta integração de técnicas *end-to-end* com DNN-HMM esteja simplificando muito a construção de modelos, estes continuam sofrendo degradação devido a ruídos, acentuações, etc.

Na tentativa de minimizar o impacto dos ruídos, o trabalho propõe um modelo que usa redes neurais adversárias assim como *Generative Adversarial Network* (GAN) e *Domain Adversarial Training* (DAT). No artigo, os autores usaram A*dversarial Training* (AT) baseado em *Fast Gradient Sign Method* (FGSM). Este modelo foi testado sobre as bases AISHELL-1 e AISHELL-2 em Mandarim.

Durante o desenvolvimento, os autores realizaram diversos testes, inclusive alterando para A*dversarial Regularization* (AR), onde obtiveram melhor resultado, com melhoria de aproximadamente 18% sobre o modelo normal LAS.

Deste artigo foi possível extrair as seguintes informações:

- Hidden Markov Models (HMMs) é um modelo tradicional para reconhecimento de fala;
- CTC e "A*ttention based Model*" são outros modelos, focados em tarefas de sequência.
- Redes neurais adversárias melhoraram o resultado da rede.

3.2.6.4. SoundSifter: Mitigating Overhearing of Continuous Listening Devices (Islam M, Islam B, Nirjon S, 2017)

O artigo é focado no desenvolvimento de um *case* junto de um *hardware* para realizar filtragens de frequência e realizar a separação de vozes e ruídos. O principal objetivo dos autores foi o desenvolvimento de uma rede neural que conseguisse distinguir bem entre sons primários e secundários. Os sons primários, segundo os autores, são os que possuem frequência maior, por se encontrarem próximos ao dispositivo de entrada e serem contínuos ao longo da amostra. Os sons secundários são os que possuem frequência mais baixa, por estarem mais distantes do dispositivo de entrada, serem intermitentes e não contínuos ao longo da amostra. Alguns exemplos de sons secundários são alarmes, por serem intermitentes, ou ainda, carros passando, por não estarem presentes em toda a amostra.

O trabalho é focado inteiramente no desenvolvimento da rede neural, junto com o *case*. Para validação foi realizada uma comparação simples entre as respostas do Amazon Echo[2] e o *case*, frente a diversos cenários com ruído. Segundo os autores, o *SoundSifter* obteve resultados melhores, conseguindo responder 46 comandos em relação a 26 da Amazon Echo.

Deste artigo foi possível extrair as seguintes informações:

- Uma possível boa abordagem para melhorar a performance de uma rede de conversão de fala em texto, é tentar extrair as frequências principais das secundárias.

3.2.7. Discussão

Apesar de no final do estudo terem sido selecionados apenas quatro estudos, estes foram de grande valia para o intuito proposto, pois as informações obtidas servirão como base para a tomada de decisão para a continuação deste TCC.

3.2.7.1. Sumário da evidência

O Quadro 5 mostra as informações que puderam ser extraídas dos estudos:

[2] *Hardware* com alto-falante, integrado a assistente virtual Alexa da Amazon. Existem vários modelos, desde simples apenas com autofalantes e captura da voz, até integrados com *display* de vídeo.

Quadro 5 – Técnicas obtidas com a revisão sistemática.

Técnica	Objetivo	Método de desenvolvimento
Hidden Markov models (HMMs) + DNN	Conversão de voz em texto	Construção própria
attention-based encoder-decoder (AED)	Conversão de voz em texto	Construção própria
Google cloud (Speech to Text)	Conversão de voz em texto	API de terceiros
Google cloud (Translate)	Tradução de uma linguagem para outra	API de terceiros
YOLOv3	Rede neural de classificação	Construção própria + biblioteca pronta da linguagem Python
Python NLTK	Rede neural para extração de informações	Construção própria + biblioteca pronta da linguagem Python
Redes neurais adversárias	Técnica para melhorar a performance da rede neural	Construção própria
Filtragem de frequências	Técnica para melhorar a performance da rede neural	Construção própria

Fonte: Elaborado pelo autor

Podemos observar que a extração se foca mais nas técnicas relacionadas à *machine learning*, isto se deve ao fato dos artigos selecionados serem mais voltados para esta área. Apenas o *StoryDrawer* desenvolveu um aplicativo, apresentando alguns de seus recursos e informando a técnica utilizada, contudo, as tecnologias escolhidas para o aplicativo propriamente dito não foram informadas.

3.2.8. Considerações finais sobre a revisão sistemática

Baseado no estudo realizado, foi possível concluir que o desenvolvimento de um conversor de voz em texto não é muito convencional, sendo complexo e exigindo muito estudo em cima das técnicas a serem utilizadas. Com isto em mente, uma possível abordagem a ser empregada para o desenvolvimento do aplicativo de finanças com comandos de voz, seria em um primeiro momento utilizar de APIs de terceiros, como foi o caso do artigo *StoryDrawer,* utilizando Google *Cloud* para a conversão de voz em texto. Em um segundo momento, caso não se desejasse manter a dependência de serviço terceirizado, construir uma rede neural com esta finalidade.

4. ARQUITETURA DO SISTEMA

Com a construção do referencial teórico e a realização da revisão sistemática foi possível obter *insights* para a tomada de decisão dos recursos a serem utilizados neste projeto. A principal definição obtida é que para entender e executar ações com base na fala, não é necessário utilizar um modelo diretamente ligado a voz. É possível primeiramente converter a voz em texto e então utilizar um modelo de interpretação textual. Esta abordagem possui mais vantagens em relação ao treinamento com base nas frequências, timbre, dentre outras características do som, pelos motivos que podem ser vistos no Quadro 6:

Quadro 6 - Vantagens e Desvantagens das abordagens de extrair informação.

Tecnologia usando diretamente a voz	Tecnologia usando texto (Voz convertida previamente)
Vantagens / Desvantagens	Vantagens / Desvantagens
V – É direto, não necessitando de um processo intermediário.	D – Como o processo é por voz, necessita de um processamento intermediado para gerar o texto a ser analisado.
V – Acaba sendo mais preciso, pois pode usar várias características (timbre, amplitude, frequência, etc.) para aprender.	D – Só tem como aprender com base nos caracteres, palavras.
D – A geração da base de treinamento necessita ser ampla contendo vozes de diversas pessoas.	V – A geração da base de treinamento é mais simples, não necessita envolver várias pessoas, pode ser ampla, usando textos da internet como base.
D – A armazenamento da base de treinamento, mesmo uma base pequena, utiliza muito espaço em disco.	V – Como a base de dados usa apenas caracteres, mesmo uma base grande, utiliza pouco espaço.
D – Pode interpretar ruídos como um comando.	D – Pode interpretar letras que não formam uma palavra como informação.

Tecnologia usando diretamente a voz	Tecnologia usando texto (Voz convertida previamente)
Vantagens / Desvantagens	Vantagens / Desvantagens
D – A indexação dos dados é complexa.	V – Mesmo inserindo informações extras para indexação, o *dataset* continua pequeno comparado ao de voz.
D – Pouco material sobre o assunto para pesquisa.	V – Muito material na internet, mas é necessário filtrar bem.

Fonte: Elaborado pelo autor

Seguindo este raciocínio, para a realização da primeira tarefa, de converter a voz em texto, foi utilizada a API *Speech to Text* da plataforma do Google *Cloud*. Esta ferramenta foi utilizada pelo artigo *StoryDrawer* (Zhang C, Yao C, Wu J et al, 2022). Para a segunda tarefa, que é o reconhecimento e extração de informações destes comandos, foi utilizada a biblioteca NLTK, a partir da linguagem Python.

4.1. Definições sobre o projeto

Apesar das diversas funcionalidades presentes nos aplicativos de gerência financeira, conforme apresentado no referencial teórico, nem todas são cabíveis ou fazem sentido de serem abordadas por comandos de voz. Para o presente trabalho as ações que foram definidas como fundamentais no aplicativo de gerência financeira foram:

- Registro de despesa: registro de débitos em uma conta/carteira, de modo a diminuir o saldo.
- Registro de receita: registro de crédito em uma conta/carteira, de modo a aumentar o saldo.

Desta forma, os comandos a serem treinados devem estar relacionados a uma destas duas ações, possuindo complicadores relacionados apenas à obtenção das informações presentes no texto, como categoria, valor, descrição, etc. O Quadro 7 demonstra exemplos de frases com as informações extraídas em cada frase.

Quadro 7 - Exemplo de classificação e informações presentes nas frases

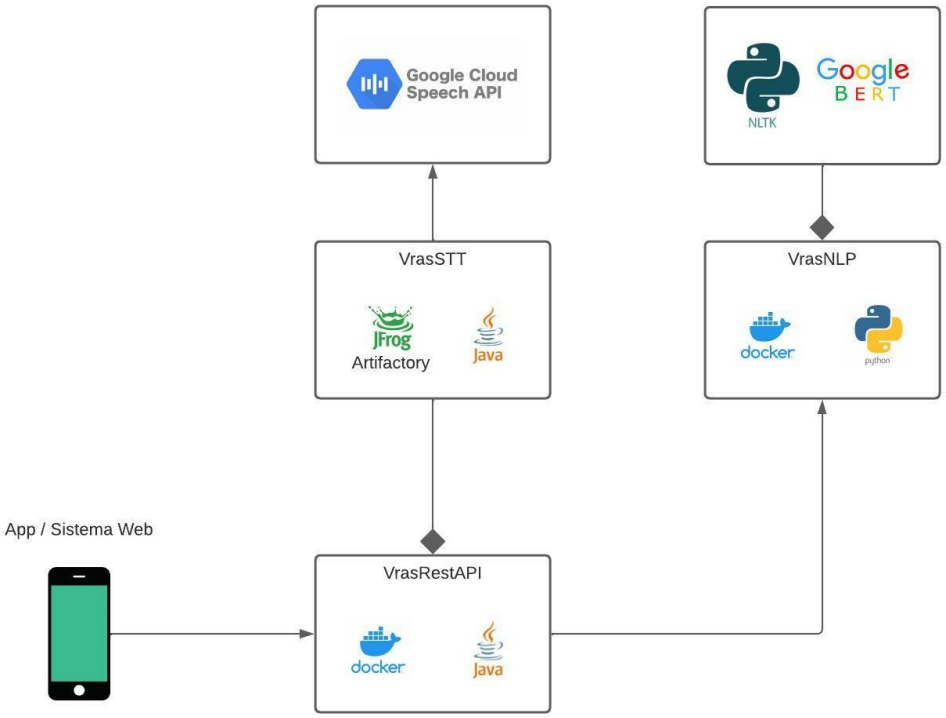
Frase	Ação	Categoria	Valor
Supermercado Max compras do mês r$ 609	Despesa	Mercado	R$ 609
paguei a fatura da Claro deu 258 com 17	Despesa	Serviços	R$ 258,17
Shopping Novo Hamburgo Lojas Renner r$ 120	Despesa	Vestuário	R$ 120
Recebimento de salário do mês 2283 com 25 centavos	Receita	Salário	R$ 2283,25

Fonte: Elaborado pelo autor

A Figura 12, demonstra o modelo da arquitetura do projeto. Nele podemos verificar quatro projetos e uma integração. A sigla “Vras” foi elaborada pelo autor com o significado: *Voice Recognition Assistant System* (VRAS), em português, Sistema Assistente de Reconhecimento de Voz. Os projetos são os seguintes:

- App / Sistema Web: interface pela qual o usuário realizará as ações de despesa, receita ou transferência.
- VrasRestAPI: projeto centralizador das requisições que designa os recursos necessários.
- VrasSTT: STT (*Speech to Text*) responsável por converter a voz em texto, este possui integração com *Google Cloud Speech to Text.*
- VrasNLP: responsável por interpretar os comandos de texto, gerando um JSON das informações obtidas, utilizando bibliotecas NLTK e BERT.

Figura 12 - Modelo macro dos componentes do projeto.

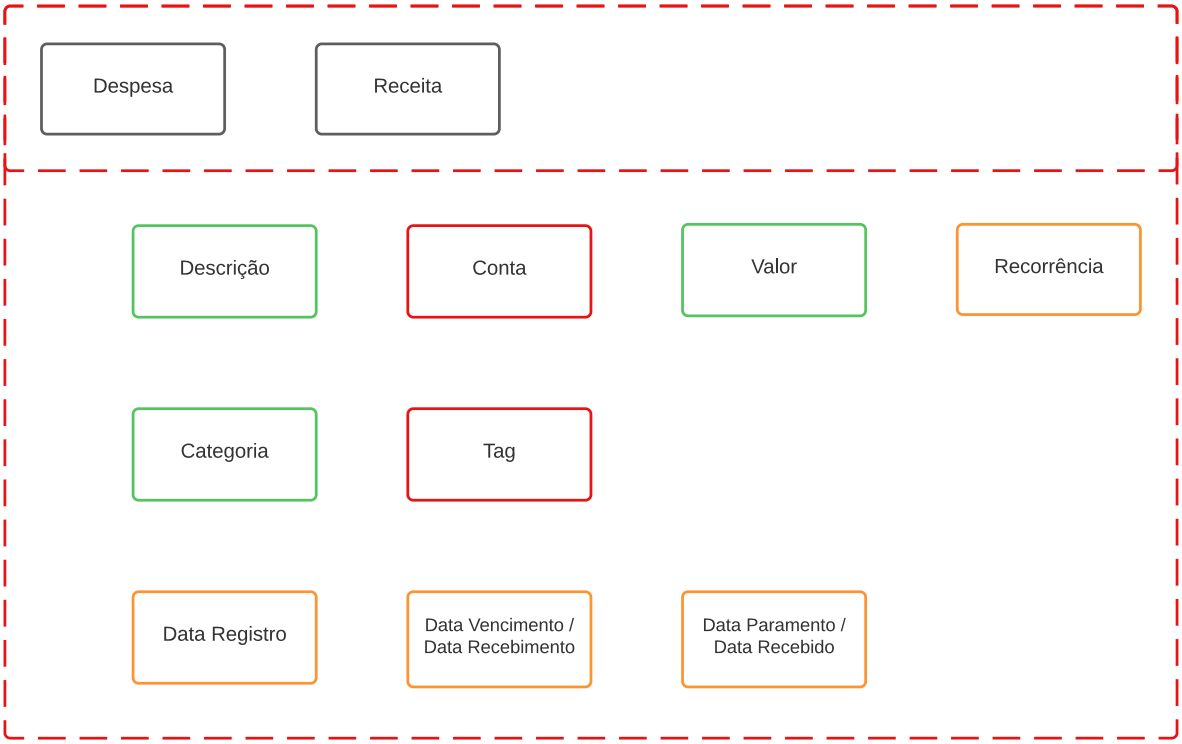


Fonte: Elaborado pelo autor

4.2. As Ações

As ações podem ser do tipo despesa e receita, cada uma delas possui suas propriedades. Uma ação de despesa, por exemplo, possui três propriedades: valor, descrição e categoria. Foi elaborado um modelo, conforme a Figura 13, na qual o retângulo pontilhado superior contém as ações relacionadas às propriedades do retângulo pontilhado inferior. As cores estão relacionadas à possibilidade de a propriedade aparecer nos comandos de voz. O verde indica que a propriedade aparece em grande frequência das vezes, o laranja com menos frequência e o vermelho indica uma propriedade pouco presente. Estas definições foram obtidas observando o *dataset* gerado, que será apresentado a seguir.

Figura 13 – Modelo contendo ações e propriedades presentes na ação.



Fonte: Elaborado pelo autor

4.3. Convertendo voz em texto (VrasSTT)

Para a tarefa de converter a voz em texto foi utilizada a API *Speech to Text* da plataforma do Google *Cloud*. Esta API possui documentação e recursos de integração para a plataforma Java. O desenvolvimento desta aplicação, denominada como VrasSTT, utilizou as seguintes tecnologias:

- Java 17
- Spring Boot
- Google Cloud - *Speech to Text*
- *Artifactory*

Com o intuito de criar uma aplicação de maior abrangência, foi definido que esta deveria tratar apenas a conversão de voz em texto. Demais ações deverão ser tomadas pelas aplicações que a utilizam.

Para a criação desta aplicação foi definida a utilização do protocolo HTTP, com o modelo REST API. Os passos para a conversão de voz em texto são os seguintes:

- É realizada uma solicitação contendo um arquivo de áudio para a conversão.
- A API verifica se o arquivo está nos padrões configurados.
- O arquivo é enviado à API do Google Cloud, que realiza a conversão.

- A resposta, em formato JSON, é formulada contendo o texto que foi reconhecido e a precisão da conversão.

4.4. Gerando o *dataset* inicial

Após a criação do VrasSTT, a fim de testá-lo, assim como gerar um *dataset* para o processo de tokenização (que será abordado a seguir), foram gravados ao total 200 áudios, criados pensando nas ações de despesa e receita, contendo diversas categorias, valores e descrições. Embora a aplicação aceite entradas de áudio, decidiu-se por utilizar o *Whatsapp* para a gravar e realizar *download* a partir do *Whatsapp Web.*

Após serem submetidos a API REST (VrasSTT), os resultados obtidos foram salvos em um arquivo, gerando assim o *dataset* inicial. O Quadro 8 demonstra algumas das convenções que foram realizadas pela API, estando no mesmo formato de saída desta.

Quadro 8 – Exemplo de frases convertidas em texto

Supermercado Max compras do mês r$ 609
Farmácia Panvel r$ 150
paguei a fatura da Claro deu 258 com 17
areia para os gatos 48 com 15
presente de aniversário para mãe 100
Shopping Novo Hamburgo Lojas Renner r$ 120
Supermercado r$ 37 com 14 centavo

Fonte: Elaborado pelo autor

4.5. Tokenização do texto

Com base na informação obtida pela revisão sistemática, verificou-se que o processo utilizado para “entender” o que está no texto é denominado tokenização. Neste processo o modelo é treinado para gerar um classificador para cada uma das palavras presente no texto. Por exemplo, classificá-las de acordo com sua estrutura gramatical: verbo, adjetivo, artigo, etc. Para o desenvolvimento deste processo utilizou-se a biblioteca NLTK da linguagem Python, descoberta na revisão sistemática como uma técnica possível.

O processo de classificação por *tokens* funciona como um conjunto finito, onde cada palavra da frase é “substituída” por um *token* deste conjunto. O objetivo da IA

neste processo, é encontrar o *token* “ideal” de cada palavra, dada a posição em que ela se encontra e o contexto da frase. – Este processo está ligado a área de Processamento da Linguagem Natural (PLN) - do inglês NLP (*Natural Language Processing*).

A biblioteca NLTK possui diversos modelos pré-treinados por terceiros; por convenção este treinamento foi denominado “*corpus*”, sendo uma coletânea ou conjunto de documentos sobre um determinado tema.

No desenvolver de um modelo de tokenização, o primeiro passo é definir o conjunto de *tokens* finitos com seus respectivos objetivos. Por exemplo, o modelo “Mac-Morpho” presente na biblioteca NLTK do Python, utiliza um conjunto de tokens para estrutura gramatical, conforme pode ser visto no Quadro 9. O segundo passo é criar um *dataset* com textos onde cada palavra do texto recebe o *token* desejado. O terceiro já é o treinamento do modelo, onde o modelo vai aprender a classificar o token correto para cada palavra, dada a posição em que ela se encontra e o contexto da frase.

Durante os testes foram realizadas 2 abordagens: a primeira utilizando o *corpus* da própria biblioteca NLTK, este *corpus* foi treinado em português e utiliza *tokens* de estrutura gramatical, conforme mencionado anteriormente. Por exemplo, o *token* (V) se refere a um Verbo, já o *token* (NPROP) se refere a um “nome próprio”. Mais exemplos podem ser vistos no Quadro 9. Os testes utilizando este modelo já treinado não foram satisfatórios.

Quadro 9 - Exemplo de tokens presentes do corpus Mac-Morpho da biblioteca NLTK.

Token	Significado
ART (Artigo)	É a classe de palavras que morfologicamente variam em gênero e número e são sempre pré - nominais, determinando o sintagma nominal de forma determinada ou indeterminada.
ADJ (Adjetivo)	Classe das palavras que geralmente funcionam como modificadores de um sintagma nominal. Podem ocorrer tanto como pré ou pós-nominais e flexionam-se em gênero e número.

N (Nome)	Classe das palavras que geralmente desempenham o papel de núcleo em um sintagma nominal
NUM (Numeral)	Fazem parte da classe dos numerais apenas aqueles chamados cardinais (grafados ou em forma de números), enquanto pré-nominais.

Fonte: Elaborado pelo autor (extraido do Manual da biblioteca Mac-Morpho)

Os comandos de áudio procuram ser mais diretos e não utilizam a estrutura gramatical completa. Por exemplo, “Xis da voni sábado à noite r$ 38 com 25 centavos”. Esta frase tem significado, pois conseguimos extrair informações dela, mas gramaticalmente falando, falta coesão textual. Por exemplo, ao comunicar, a uma pessoa, a mesma frase, provavelmente seria “Fui ao Xis da Voni no sábado à noite e paguei R$ 38 com 25 centavos”.

Quadro 10 – Classificação de frases usando Corpus Mac Morpho

Índice	Frase onde as palavras possuem o token relacionado	Situação do token
1	paguei/V a/ART fatura/N da/N Claro/ADJ deu/V 258/NUM com/PREP 17/NUM	Correto
2	remédio/N da/N pressão/N duas/NUM caixas/N 39/N|AP com/PREP 35/NUM	Incorreto
3	presente/ADJ de/PREP aniversário/N para/PREP mãe/N 100/N	Incorreto
4	Supermercado/N Max/NPROP compras/N do/KS mês/N r/N $/$ 609/N com/PREP 27/N centavos/N	Incorreto

Fonte: Elaborado pelo autor

O principal problema estava relacionado ao valor, pois a sentença com esta informação deveria estar relacionada a *tokens* de numeração (NUM), conforme item 1 do Quadro 10. Contudo, por várias vezes os valores eram classificados como “nome” (N), assim como pode ser observado no item 3 do Quadro 10, em que o “100” foi atribuído a “mãe”, como se estivesse compondo o nome de uma pessoa (“Dom Pedro I”), local (“25 de março”), instituição (“Colégio Franciscano Pio XII”), etc. Outro problema estava na classificação de outras sentenças, não relacionadas ao valor, que

também recebiam a classificação de nome (N), conforme item 4 do Quadro 10, o que dificultou o processo de identificação do valor.

Para tentar obter uma classificação mais assertiva para os *tokens* de valor, assim como classificar a frase, decidiu-se criar um *corpus* pessoal, onde informações numéricas sempre são classificadas com *token* numérico (NUM). Na sequência, também foi criado um novo *token* (ACT) para identificar palavras que tenham relação com uma das duas ações desejadas, por exemplo “paguei” está relacionado a ação despesa, “salário” está relacionado a ação receita. Desta forma estas palavras recebem o *token* ACT para, em processo posterior, seja possível identificar qual ação deve ser tomada.

Para o desenvolvimento deste modelo, foi necessária a criação de um *corpus* personalizado, com frases contendo os tokens desejados. Desta forma um *dataset* foi gerado manualmente com auxílio do recurso VrasSTT. Esta abordagem mostrou-se efetiva para a identificação do valor, mas ainda não resolveu o problema de classificar a frase em despesa ou receita, conseguindo apenas encontrar as palavras que pudessem identificar a ação. Outro problema está relacionado ao esforço para a criação do *corpus*, onde além de criar as frases, era necessário classificar cada uma de suas palavras com o *token* desejado.

4.6. Alternativas à tokenização

Mediante aos problemas para classificar as frases corretamente com a abordagem de *tokens* pelo NLTK, iniciou-se uma pesquisa para realizar esta tarefa. Baseando-se na revisão sistemática, realizou-se diversas pesquisas para classificação de frases ou texto, usando técnicas de *tokens* ou mesmo similares ao modelo de classificação YOLOv3. Dentre estas pesquisas a mais promissora foi a relacionada a “redes neurais para análise de sentimentos”, que será abordada no tópico a seguir. Esta abordagem diz respeito a classificação de frases para obtenção de *insights* sobre como o público está reagindo a um determinado produto ou serviço.

Ainda para obtenção de um método mais assertivo, foi realizada uma reunião com o professor da Universidade Feevale, Gabriel da Silva Simões, doutor na área de inteligência artificial. Ele afirmou que de fato este era um problema de classificação, contudo sendo de classificação múltipla e a análise de sentimentos seria uma boa abordagem para esta tarefa, recomendando a utilização da IA BERT do Google, um dos mais recentes modelos utilizados para estas tarefas.

Coincidentemente, neste mesmo período ocorreu a palestra no evento BARNLP[3], na Universidade Feevale, com o tema sobre IA e os avanços da área de PLN, na qual um dos palestrantes falou sobre análise de sentimentos e a IA BERT, reafirmando as pesquisas realizadas, assim como as recomendações do professor.

4.6.1. Redes neurais para análise de sentimentos

Segundo Pedro C T Gomes 2019, os profissionais que trabalham com ciência de dados, desenvolveram diversas ferramentas para gerar informações e *insights* sobre grande quantidade de dados coletados pelas empresas. Uma destas ferramentas é a análise de sentimentos, que visa entender como o público está reagindo a determinado produto ou serviço fornecido pela empresa, “ela ajuda as empresas a entenderem o sentimento social de sua marca”. Através da coleta de frases e avaliações deixadas nas mídias sociais, páginas, etc. é possível classificar sentenças, como positivas, negativas ou neutras, auxiliando inclusive a identificar oportunidades de mercado.

“Um sistema de análise de sentimentos para conteúdo textual combina o processamento de linguagem natural (PLN) e técnicas de aprendizado de máquina para atribuir pontuações ponderadas de sentimento às sentenças” (Pedro C T Gomes, 2021). Esta técnica funciona como um classificador para a sentença como um todo.

4.6.2. Google BERT

Segundo Data Science Academy (2022), o Bidirectional Encoder Representations from Transformers (BERT) foi criado em 2018, sendo um modelo pré-treinado para PLN, que utiliza algoritmo de aprendizado profundo (*Deep Learning*). Seu código é aberto, podendo ser usado por qualquer pessoa para treinar seus sistemas de processamento de linguagem.

O algoritmo do BERT, criado pelo Google, foi implantado em 2019 em seu buscador, com o objetivo de melhorar seus resultados e permitir que as pesquisas sejam realizadas de forma mais natural. Sendo o primeiro modelo bidirecional, onde “a bidirecionalidade é obtida pelo fato de as frases serem processadas tanto em relação aos dados à esquerda quanto à direita”, tornando a compreensão do texto

[3] BARNLP foi um evento gratuito com os temas: Inteligência Artificial, Processamento de Linguagem Natural e Aplicações de I.A. Neste evento foram realizadas palestra por doutores na área de tecnologia ou correlatas.

muito mais assertiva, já que considera ambos os contextos. O modelo inspirou a criação de diversos outros, como, ROBERTA, ALBERT, DistilBERT, etc. (Paiva S E, Pereira S F, 2022).

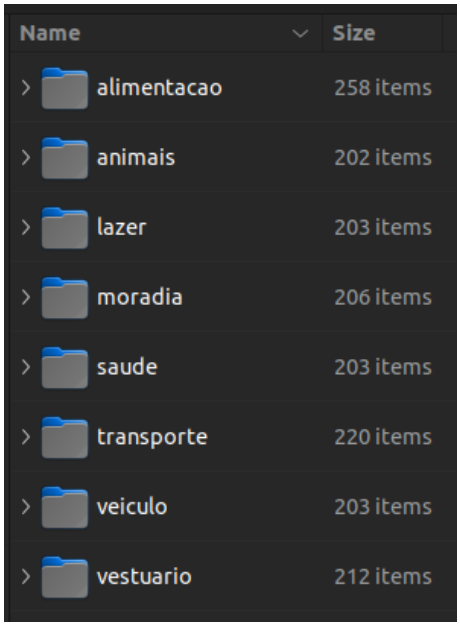
4.7. Definição da técnica de classificação

Com base na pesquisa prévia sobre análise de sentimentos e sugestão do modelo BERT, iniciou-se uma pesquisa, focada em: “classificação múltipla utilizando modelo BERT” e "Análise de sentimentos com BERT”. A pesquisa revelou diversos artigos e blogs sobre o assunto. Contudo, muitas das abordagens eram rasas, sendo de classificação simples ou com implementações complexas demais para uma classificação de 2 níveis. O artigo que se destacou, mais adequado ao trabalho em questão, foi o “Multi-label Text Classification using Transformers(BERT)” (Prasad Nageshkar, 2021), que realiza a classificação múltipla através de *tags*, não sendo tão complexo e por trabalhar com *tags*, torna-se flexível para criar quantos níveis forem necessários.

O artigo deu base para o desenvolvimento de testes de classificação múltipla focados no aplicativo financeiro. Estes foram realizados pela plataforma Google *Collaborate,* sem qualquer necessidade de instalações em máquina pessoal.

O primeiro passo foi criar um novo *dataset*. Para o desenvolvimento, utilizou-se a mesma abordagem da geração do *dataset* com *tokens*, contudo para ser mais objetivo, criou-se pastas das categorias desejadas, conforme Figura 14.

Figura 14 - Pastas classificadas para despesa



Fonte: Elaborado pelo autor

De todos os áudios gerados, foram escolhidos 200 para cada uma das categorias. O objetivo foi criar um *dataset* com diversos exemplos e ao mesmo tempo balanceado. Ao final o *dataset* possui 8 categorias, totalizando 1600 áudios. Estes foram enviados via *Whatsapp* e a conversão para texto foi através da aplicação VrasSTT.

O *dataset* para classificação possui 2 colunas, a primeira com a frase de treinamento e a segunda com as *tags* desejadas para esta frase. As *tags* por sua vez estão relacionadas a ação e categoria. O Quadro 11 demonstra detalhes sobre o *dataset*.

Quadro 11 – Exemplo de classificação de frases

Frases	*Tags*
comida para o peixe r$ 30	[despesa, animais]
peguei um Uber até o hospital 2042	[despesa, transporte]
lavadora lava e seca r$ 1200	[despesa, moradia]
gasolina aditivada r$ 100	[despesa, veiculo]
recebimento do salaria r$ 3748	[receita, salario]
recebi meu auxilio home office 150	[receita, auxilio]

Fonte: Elaborado pelo autor

Para criação do modelo de classificação foi definido que 90% do dataset seria destinado ao treinamento, totalizando 1440 linhas, enquanto 10%, foi destinado a validação do modelo, totalizando 160 linhas.

Os testes com esta abordagem se mostraram efetivos, pois com o treinamento de apenas 19 épocas (Cada época é um treinamento com pesos diferentes, onde cada época subsequente os pesos são alterados usando uma taxa de correção em relação a época anterior), foi obtido uma acurácia de 98%, precisão e *recall* acima de 95%. A Tabela 12, demonstra os resultados do treinamento de testes para despesas.

Tabela 12 – Resultados para do classificador de despesas

	precision	recall	f1-score	support
0	0.99	0.99	0.99	1120
1	0.97	0.97	0.97	320
Accuracy			0.98	1440
Macro avg	0.98	0.98	0.98	1440
Weighted avg	0.98	0.98	0.98	1440

Fonte: Elaborado pelo autor

Em sequência foi realizado o mesmo procedimento para a ação de receita, utilizando as categorias: auxílio, estorno, jogo, presente, salário, serviço, vale e venda. A Tabela 13, demonstra que os resultados foram similares aos da ação de despesas.

Tabela 13 – Resultados para do classificador de receitas

	precision	recall	f1-score	support
0	0.99	0.99	0.99	1120
1	0.97	0.97	0.97	320
Accuracy			0.98	1440
Macro avg	0.98	0.98	0.98	1440
Weighted avg	0.98	0.98	0.98	1440

Fonte: Elaborado pelo autor

Por final, para o modelo a ser utilizado na aplicação financeira, foi treinado o modelo que une ambos os *datasets* de despesa e receita. Os resultados deste podem ser vistos na Tabela 14, embora este modelo tenha obtido resultados altos, foram necessárias 38 épocas para chegar a este resultado.

Tabela 14 – Resultados para do classificador combinado

	precision	recall	f1-score	support
0	0.99	1.00	0.99	5120
1	0.97	0.94	0.95	640
Accuracy			0.99	5760
Macro avg	0.98	0.97	0.97	5760
Weighted avg	0.98	0.99	0.99	5760

Fonte: Elaborado pelo autor

4.7.1. União das técnicas Tokenização e Classificador BERT

Apesar do resultado obtido pelo sistema de classificação, utilizando BERT, para concluir o processo desejado na aplicação VrasNLP, ainda era necessário extrair as informações de descrição e valor presentes na frase, contudo o processo de tokenização havia obtido bons resultados no processo de extrair estas informações. Desta forma decidiu-se unir ambas as técnicas.

4.8. Identificação das características presentes na frase (VrasNLP)

O objetivo da aplicação VrasNLP é, conforme mencionado anteriormente, “interpretar os comandos de texto, gerando um JSON das informações obtidas”, ou seja, com base em uma frase enviada a ele, sua resposta deve ser, conforme Quadro 7, um JSON informando qual ação deve ser tomada para realizar o registro, assim como, categoria, descrição e valor relacionado à frase.

O desenvolvimento desta aplicação, utilizou as seguintes tecnologias:

- Python 3.8
- Docker
- Python - Biblioteca NLTK
- Python - Algoritmo BERT

A obtenção da ação e categoria da frase foi realizada utilizando a técnica de classificação, apresentada na seção 4.7. A descrição e valor foram obtidos usando a tokenização apresentada na seção 4.5.

Com a junção das técnicas mencionadas, a aplicação VrasNLP é capaz de receber uma frase e devolver um JSON das informações que foram extraídas. Em caso de informação incompleta, como no caso do usuário esquecer de informar o

valor, o JSON continua enviando as demais informações, mas com valor “null”, ficando o usuário responsável por informar os dados faltantes na consolidação do registro.

4.9. Gerenciador unido de processos (VrasRestAPI)

As aplicações VrasSTT e VrasNLP cumpriram seus objetivos específicos, contudo estão separadas e apenas geram informações baseadas em suas entradas, respectivamente áudio e texto. Ainda é necessário realizar validações sobre as informações, tanto as de entrada quanto as de saída. É importante ressaltar que, embora já tenhamos boa parte das informações desejadas, ainda não foi criado de fato um registro, é necessário consolidar as informações geradas em um banco de dados e criar um registro a ser valido pelo o usuário.

Com este objetivo em mente foi desenvolvida a aplicação VrasRestAPI, que centraliza ambas as aplicações mencionadas anteriormente e ainda gera um registro a ser validado. Na pratica, foi criado um *endpoint* que recebe o áudio, extrair as informações, consolidar no banco de dados como um registro e, ao final do processo, retorna o mesmo para ser feita a validação pelo usuário.

Embora em uma visão externa pareça que a aplicação extraiu as informações do áudio, o que ocorreu internamente foi a utilização da aplicação VrasSTT para converter o áudio em texto e o resultado foi fornecido como entrada para a aplicação VrasNLP que extraiu as informações, por último a própria aplicação VrasRestAPI, ficou responsável por consolidar as informações extraídas no banco de dados.

A seguir no Capítulo 5, serão descritos demais processos envolvidos na aplicação da VrasRestAPI.

4.10. Considerações finais sobre a Arquitetura

Podemos verificar que o processo de criação de uma aplicação com recursos de voz, é bastante amplo. É possível obter *insights* sobre frases e textos utilizando apenas uma rede neural para classificação, assim como demonstra a análise de sentimentos e o teste realizado com algoritmo BERT. Contudo, caso o objetivo da aplicação envolva extração de dados específicos, será necessário utilizar técnicas de *Information Extract* (IE), conforme demonstrado ao utilizar a Tokenização.

Todo o desenvolvimento envolvido neste capítulo foi essencial para o desenvolvimento da aplicação desejada, sendo o cerne do sistema.

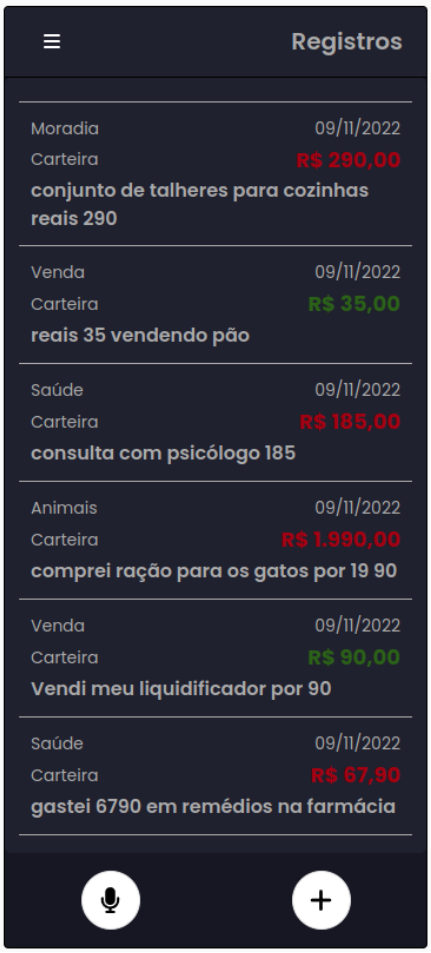
Os testes e o desenvolvimento realizado com a biblioteca NLTK foram efetivos para a solução da extração das informações presentes nas frases. Estes também forneceram *insights* para melhorias e expansão de recursos.

5. PROCESSO DE VALIDAÇÃO DA APLICAÇÃO

No Capítulo 4 foi apresentado o processo de criação dos principais recursos para alcançar o objetivo pretendido na elaboração deste trabalho: “a compreensão de um comando de voz, sendo capaz de realizar um registro financeiro”.

Para o processo de validação dos registros, foi desenvolvida uma aplicação básica, simulando um sistema financeiro, conforme Figura 15.

Figura 15 - *Frontend* criado para testes reais



Fonte: Elaborado pelo autor

Esta aplicação conta com: sistema de usuário, sistema de contas com balanço financeiro[4], sistema para registro dos três tipos de lançamentos mencionados neste trabalho (receita, despesa e transferência) e um sistema de preferências, no qual o usuário deve configurar características em seu controle financeiro.

[4] Demonstrativo contábil em que se confrontam num dado momento, as receitas e despesas orçamentárias.

Todos estes recursos foram desenvolvidos sob a aplicação VrasRestAPI, gerando *endpoints* que são utilizados pela aplicação *frontend*. Em um primeiro momento foi utilizado o *software* Postman[5] a fim de realizar as solicitações e testes. Posteriormente, criou-se um *frontend* personalizado para testes mais realistas da aplicação financeira.

5.1. Desenvolvimento do sistema financeiro

Ao final do capítulo 4, foi possível consolidar em banco de dados, um registro criado através da extração de informações de um comando de voz. Embora o registro já esteja sendo feito, ele ainda necessita da validação do usuário, confirmando o que foi "entendido", ou corrigindo algum detalhe, como valor, categoria ou descrição.

Em um sistema financeiro se faz necessário uma série de recursos. O registro de receita e despesa é apenas um deles. Também são necessários processos de manipulação de informações e das transações, além de um demonstrativo da "saúde financeira" do usuário.

A seguir, serão discutidos alguns dos recursos considerados necessários para a validação desta aplicação financeira.

5.1.1. Sistema de gerência de usuário e *token* de login

O sistema de gerência do usuário foi desenvolvido juntamente com um sistema de "*token* de login". Neste sistema um *token* é gerado e vinculado a um determinado usuário. Este *token*, posteriormente é enviado nas solicitações ao sistema, que se encarrega de validar e verificar o usuário a qual o *token* está vinculado. Nesta abordagem, as informações do usuário não estão envolvidas diretamente.

Neste momento, por se tratar de uma validação de sistema, foi criado um usuário padrão e fixo, denominado "Root". Neste momento não será criado o sistema de login e, portanto, foi gerado um *token* randômico, vinculado manualmente a este usuário. Como mencionado, o sistema se encarrega de entender que o *token* pertence ao usuário "Root".

[5] Postman é uma ferramenta que dá suporte à execução de testes de APIs e requisições em geral. Com este *software* é possível construir solicitações rapidamente e, ainda, guardá-las para uso posterior, além de conseguir analisar as respostas enviadas pela API.

5.1.2. Categorias Padrões e pertencentes ao usuário

Diversas categorias consideradas padrões foram criadas, com o objetivo de facilitar a utilização inicial por parte dos usuários. Algumas delas são: Saúde, Alimentação, Veículo, Vestuário, Animais, Moradia, etc.

Ao realizar o cadastro de um novo usuário, as categorias padrões são copiadas para ele, passando a pertencerem a este usuário. Desta forma cada usuário do sistema pode personalizar suas categorias, criando novas, editando ou excluindo, sem alterar as categorias padrões do sistema.

5.1.3. Contas e sistema de saldo

Da mesma forma que as categorias foram geradas, cinco contas consideradas padrões para foram criadas um usuário, são elas: Carteira Pessoal, Conta Corrente, Conta Salário, Poupança e Cartão de Crédito. Sempre que um usuário é criado, estas cinco contas são copiadas para ele. Assim como em categorias, o usuário pode criar, editar ou excluir suas contas, tudo sem alterar as contas padrões.

Uma conta é um centralizador das transações financeiras, relacionada a um determinado meio no qual o usuário manipula seu dinheiro. Por exemplo, a Carteira Pessoal é um meio físico onde o usuário manipula seu dinheiro utilizando as cédulas, já a Conta Corrente, está relacionada a sua conta bancária, na qual o usuário manipula seu dinheiro digital.

Sempre que é realizado um registro, seja de receita, despesa ou transferência, este deve estar relacionado a uma de suas contas, para que seja atualizado o demonstrativo de transações financeiras nesta conta e ao final seja exibido o montante de dinheiro que o usuário possui, ou seja, seu saldo.

5.1.4. Definições do usuário

A definição de conta a qual o registro pertence não foi abordada pelos comandos de voz neste trabalho. Ficando desde já, como uma melhoria futura aos comandos a serem disponibilizados no aplicativo financeiro.

Como no momento esta definição é importante para a efetivação do registro, foi criado um sistema de definições relacionadas ao usuário. Cada usuário no sistema terá suas próprias definições, servindo para auxiliar na tomada de decisão sobre os recursos disponibilizados no sistema financeiro. O parâmetro “conta padrão” foi introduzido no sistema com o objetivo de definir a conta a qual o registro deve ser

vinculado, caso o comando de voz não consiga identificar a extração das informações. No momento este parâmetro está sendo utilizado sempre, contudo quando os comandos de voz forem aperfeiçoados, seu uso será sob demanda.

5.1.5. Registro de despesa e receita

Seguindo as definições realizadas no tópico 4.1, o primeiro passo para criar o sistema de registro foi definir os tipos de lançamentos, sendo eles, receita e despesa.

Também foi definido que um registro deve estar sempre vinculado a um usuário, assim como a uma conta, para que seu valor tenha impacto direto sobre o saldo desta conta, atualizando assim o demonstrativo geral do usuário.

O sistema de registro foi desenvolvido para ser independente. Logo, é possível que o usuário crie um registro sem a necessidade de utilizar somente o comando de voz, sendo este um facilitador para não necessitando informar tudo manualmente.

O sistema de voz foi integrado ao sistema de registro utilizando comparação. Como as informações de tipo de registro e categoria, extraídas na aplicação VrasNLP, são independentes das configurações feitas pelo usuário no sistema financeiro, o meio utilizado para criar o registro de modo mais similar às configurações do usuário, foi a comparação direta. Quando o retorno do comando de voz indicar a categoria “Farmácia”, por exemplo, é realizada uma busca dentro das categorias do usuário pela categoria que tenha nome similar, como por exemplo “farmácia”, criando assim o vínculo com uma categoria pertencente ao usuário. No caso do sistema não encontrar uma categoria similar à indicada pela extração do comando de voz, automaticamente ela é aplicada à categoria “Outros”, padrão para todos os usuários do sistema.

5.2. *Frontend* para testes realistas

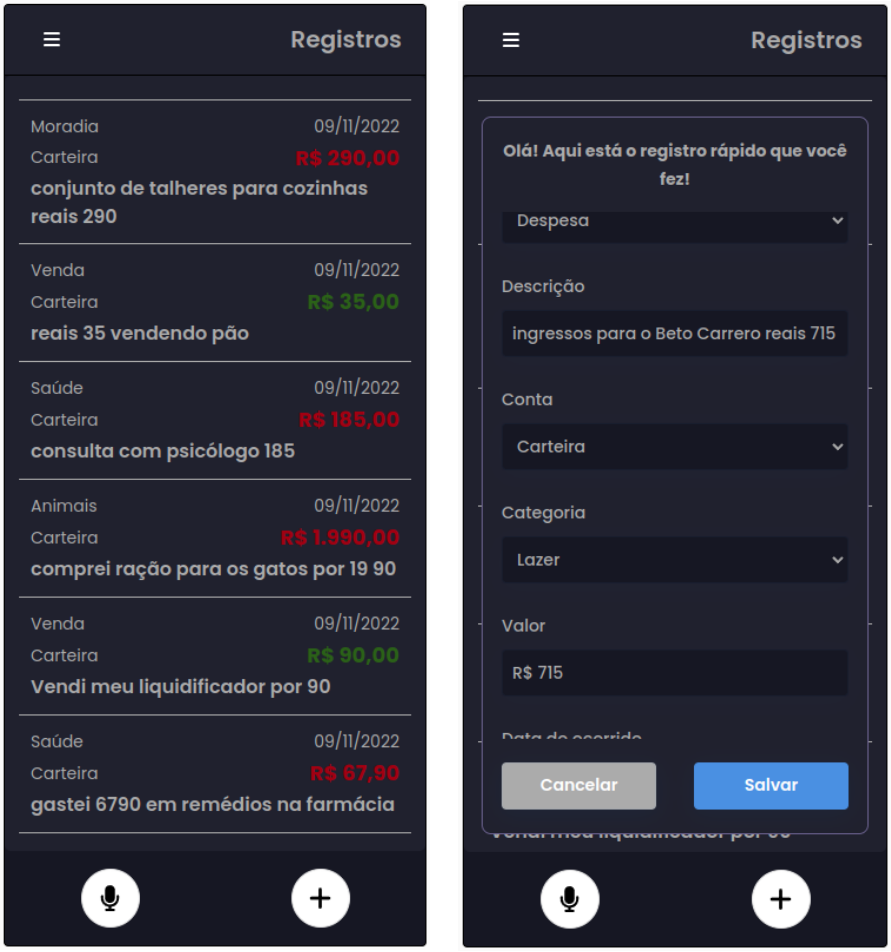
Como mencionado no início deste capítulo, em um primeiro momento, para realização dos testes do desenvolvimento relatado na seção 5.1, foi utilizado o *software* Postman. Entretanto, para testes mais realistas e que possam simular a experiência do usuário, foi desenvolvida uma aplicação *frontend* utilizando *React*[6].

Em um primeiro momento foi desenvolvida a tela com as funcionalidades de exibição dos registros efetuados e cadastro de novo registro. Nesta tela foi incluído o

[6] React é um *framework* para desenvolvimento de aplicações *frontend*, relacionadas a *web browsers.*

botão de registro por voz, com ícone de microfone, conforme pode ser visto na parte inferior da Figura 16.

Figura 16 - Tela desenvolvida para realizar os registros



Fonte: Elaborado pelo autor

Nos testes durante o desenvolvimento desta tela, percebeu-se algumas peculiaridades que podem ocorrer no cenário de uso real:

1. O usuário desiste de realizar o áudio no meio da gravação.
2. O usuário pressiona o botão de comando por áudio, mas não fala nada, gerando assim um áudio “vazio”.
3. O botão de envio é pressionado, sem querer, antes do fim da frase, gerando um áudio incompleto.
4. Excesso de barulho de fundo no momento de envio, o que inviabiliza uma identificação assertiva.

Embora estes problemas possam ocorrer naturalmente na utilização do aplicativo, nenhum deles é de fato um problema, necessitando o usuário apenas descartar o envio atual e realizar novo envio de áudio.

5.3. Validação geral, uso da aplicação financeira criada

Após o desenvolvimento da aplicação financeira que é composta pelas seções 4, 5.1 e 5.2, realizou-se a simulação real. Primeiramente foi definido que os áudios para teste, devem estar limitados às categorias que foram abordadas no treinamento de receita e despesa. Frases como “Comprei uma maquiagem nova por R$ 80,75” ou “Comprei um condicionador por R$ 18” pertencente às categorias “estética” e “higiene” respectivamente, não puderam ser utilizadas, tendo em vista que provavelmente seriam categorizadas como “Outros”, já que não estavam dentre as categorias treinadas. Ficando desde já a adição de novas categorias como uma melhoria futura ao aplicativo financeiro.

Para realizar o teste final da aplicação foram criados 100 registros, mistos entre receita e despesa e qualquer uma das categorias treinadas. Os Quadros 15 e 16, demonstram os detalhes da distribuição realizada.

Quadro 15 - Distribuição dos registros de teste real

Tipo de registro	Total
Despesa	67
Receita	33

Fonte: Elaborado pelo autor

Quadro 16 - Distribuição dos 100 registros de teste real por categoria

Tipo de registro	Categoria	Total
Despesa	Alimentação	14
	Animais	5
	Lazer	6
	Moradia	14
	Saúde	6
	Transporte	8
	Veículo	6
	Vestuário	8
Total		67
Receita	Auxilio	3
	Estorno	6
	Jogos	7
	Presente	2
	Salário	1
	Serviço	3
	Vale	2
	Venda	9
Total		33

Fonte: Elaborado pelo autor

Durante o teste foi realizada a conferência, tanto da conversão do áudio em texto, por parte de aplicação VrasSTT, como da classificação feita pelo serviço VrasNLP, os resultados podem ser vistos nos Quadros 17, 18, 19 e 20.

Quadro 17 - Distribuição do teste dos 100 registros por assertividade

Situação do registro	Nº de registros
Correto	72
Incorreto	28

Fonte: Elaborado pelo autor

Os 28 registros com erro foram classificados com o motivo do erro. Para a classificação de “Texto incorreto”, é necessário apenas que as palavras ditas tenham sido convertidas para outras diferentes, como nos casos de “comprei um quindim por R$ 5” que foi convertido como “comprei um 15 por R$ 5” e “remédios para gripe 39 com 70” que foi convertido como “remédios para gripe 39 Com certeza”. Os Quadros 18, 19 e 20 possuem mais detalhes sobre a classificação feita.

Quadro 18 – Distribuição dos 28 erros por texto e classificação

Nº de registros	Conversão do texto	Classificação
16	Correto	Incorreta
6	Incorreto	Incorreta
6	Incorreto	Correta

Fonte: Elaborado pelo autor

Quadro 19 - Distribuição dos 28 erros por categoria

Tipo de registro	Categoria	Total
Despesa	Alimentação	6
	Lazer	4
	Moradia	6
	Saúde	4
	Veículo	1
	Vestuário	4
Total		25
Receita	Jogos	1
	Serviços	1
	Venda	1
Total		3

Fonte: Elaborado pelo autor

Quadro 20 - Distribuição dos 28 erros por categoria e motivo do erro

Tipo	Categoria	Texto	Classificação	Nº Registros	Total
Despesa	Alimentação	Correto	Incorreta	2	6
		Incorreto	Incorreta	3	
		Incorreto	Correta	1	
	Lazer	Correto	Incorreta	3	4
		Incorreto	Incorreta	1	
	Moradia	Correto	Incorreta	6	6
	Saúde	Correto	Incorreta	2	4
		Incorreto	Incorreta	1	
		Incorreto	Correta	1	
	Veículo	Incorreto	Correto	1	1
	Vestuário	Correto	Incorreta	4	4
Total				25	25
Receita	Jogos	Incorreto	Correta	1	1
	Serviços	Incorreto	Correta	1	1
	Venda	Incorreto	Incorreta	1	1
Total				3	3

Fonte: Elaborado pelo autor

Pode-se verificar que embora os resultados apresentados pelo treinamento do modelo de classificação, apresentado da seção 4.7, estejam acima de 95% de assertividade, o teste real apresentou uma taxa de acerto próxima a 70%. Se considerarmos que durante o treinamento as frases foram verificadas, ou seja, não ocorreram erros por parte do conversos de voz em texto. Podemos considerar a taxa de acerto do classificado de 72 em 88 registros (~81%).

Durante a realização do teste real, 46 frases foram usadas de forma similar ao treinamento do modelo. Destas, 9 apresentaram erro, aproximadamente 20%. Já nas frases não similares 19 em 54 apresentaram erro, aproximadamente 35%. Os resultados podem ser observados nos Quadros Quadro 21, Quadro 22 e Quadro 23.

Quadro 21 - Distribuição dos registros por similaridade

	Similares	Não Similares
Corretos	37	35
Incorretos	9	19
Total	46	54

Fonte: Elaborado pelo autor

Quadro 22 - Distribuição de similares com erro por categoria e motivo

Tipo	Categoria	Texto	Classificação	Nº Registros	Total
Despesa	Alimentação	Correto	Incorreta	1	2
		Incorreto	Incorreta	1	
	Lazer	Correto	Incorreta	2	2
	Moradia	Correto	Incorreta	2	2
	Veículo	Incorreto	Correta	1	1
	Vestuário	Correto	Incorreta	2	2
Total				9	9

Fonte: Elaborado pelo autor

Quadro 23 - Distribuição de não similares com erro por categoria e motivo

Tipo	Categoria	Texto	Classificação	Nº Registros	Total
Despesa	Alimentação	Correto	Incorreta	1	4
		Incorreto	Incorreta	2	
		Incorreto	Correta	1	
	Lazer	Correto	Incorreta	1	2
		Incorreto	Incorreta	1	
	Moradia	Correto	Incorreta	4	4
	Saúde	Correto	Incorreta	2	4
		Incorreto	Incorreta	1	
		Incorreto	Correta	1	
	Vestuário	Correto	Incorreta	2	2
Total				16	16
Receita	Jogos	Incorreto	Correta	1	1
	Serviços	Incorreto	Correta	1	1
	Venda	Incorreto	Incorreta	1	1
Total				3	3

Fonte: Elaborado pelo autor

Nos quadros Quadro 22 e Quadro 23 podemos observar que independente das frases serem similares ou não similares a maior parte dos erros ocorre na classificação, sendo 7/9 e 10/19 respectivamente, ou seja 60%. Os fatores que possuem impacto direto neste resultado são: o baixo número de frases disponibilizadas para o treinamento; e a similaridade de termos presentes em frases em categorias diferentes, ambos os casos contribuem para que o modelo não consiga encontrar alguma característica para distinguir e realizar a classificação correta.

Embora o maior número de registros tenha sido feito utilizando a ação de despesa (67), é importante verificar que dos 28 erros, apenas 3 foram de receita e mesmo entre estes 3, apenas 1 foi de fato classificado incorretamente. Podemos concluir que a classificação de receita tende a ser mais precisa que a de despesa.

5.4. Considerações finais sobre o aplicativo financeiro desenvolvido

Os testes reais na aplicação desenvolvida, demonstraram que os resultados apresentados pelo modelo de classificação não são ideais. Ficando limitados aos cenários de treinamento, mesmo que generalizem bem para classificar as frases do teste, não conseguem contemplar a ampla variedade de sentenças que compreendem os comandos de voz com a mesma finalidade. Por exemplo: “Comprei batatas na feira por R$ 15”, “Peguei batatas na feira por R$ 15”, “Fui na feira e adquiri batatas por R$ 15”, “Adquiri batatas na feira por R$ 15”, etc.

6. CONSIDERAÇÕES FINAIS

O presente estudo apresenta o desenvolvimento de uma aplicação com comandos de voz, utilizando técnicas para conversão de voz em texto, classificação e extração de informações. Foram utilizados, respectivamente, serviço do Google *cloud Speech-to-Text, a*lgoritmo BERT para classificação bidirecional e Tokenização com biblioteca NTLK da linguagem de programação Pyhton.

A partir do referencial teórico e da revisão sistemática foi possível evidenciar que aplicativos com comando de voz ainda são pouco explorados, tendo em vista o grande potencial de aplicação. Foram encontradas utilizações simples de conversão de voz em texto para informar alguma descrição no registro.

O estudo demonstra não ser necessário a criação de modelos de aprendizado por voz para poder trabalhar com comandos de voz. É possível usufruir da conversão de voz em texto por serviço de terceiros e com base no resultado, extrair informações para realizar a ação desejada. O Quadro 6 da Seção 4 discute sobre as vantagens e desvantagens destas abordagens e demonstra que caso o foco esteja no conteúdo da fala e não nas características presentes na frequência ou timbre da voz, a utilização por conversão de voz em texto é mais vantajosa e de baixo custo.

Foi realizada a classificação de frases através do modelo BERT, demonstrando ser possível obter informações sem a utilização de algoritmos de extração.

Os resultados do desenvolvimento da aplicação financeira com recursos de voz mostraram-se satisfatórios. Com taxa de assertividade de 70%, foi disponibilizado ao usuário, através do comando de voz, todo o registro que ele necessita fazer de forma manual. Foi contemplado com o comando de voz: ação do registro (receita ou despesa), categoria, valor e descrição.

O presente estudo ainda demonstra que mesmo um treinamento de modelo pequeno, onde foram utilizados 3200 registros para 16 classes, ou seja, 200 registros por classe, já foi possível alcançar uma taxa de 70% de assertividade.

Ficando desde já como melhorias futuras o aprimoramento do modelo, com mais classes para fornecer maior diversidade de categorias e ações, assim como, o número de frases para treinamento, aumentando assim a taxa de assertividade.

Para trabalhos futuros, por parte do aplicativo de gerência financeira, pretende-se alterar para linguagem de programação Flutter, fornecendo recursos nativos dos sistemas operacionais Android e IOS, tais como GPS, banco de dados interno,

registros offline, sistema de lembretes e notificações. Por parte dos comandos de voz, pretende-se implementar a adição de mais categorias conforme mencionado anteriormente, um sistema de *tags*, ação de transferência, alteração de dados, determinar se uma conta já está paga, número de parcelas e lembrete de pagamento.

REFERÊNCIAS BIBLIOGRÁFICAS

AUGUSTO C, PACHECO R, PEREIRA N. **Deep Learning Conceitos e Utilização nas Diversas Áreas do Conhecimento**. Disponivel em: <http://anais.unievangelica.edu.br/index.php/adalovelace/article/view/4132/2770>. Acesso em: 15 de Dezembro de 2022.

BACEN. **O brasileiro e sua relação com o dinheiro**. Disponível em: <https://www.bcb.gov.br/conteudo/home-ptbr/TextosApresentacoes/Apresentacao_Pesquisa_Mecir_Brasileiro_Relacao_com_Dinheiro_19072018.pdf>. Acesso em: 7 de maio de 2022.

BACEN. **Estatísticas de Meios de Pagamentos**. Disponível em: <https://www.bcb.gov.br/estatisticas/spbadendos?ano=2020>. Acesso em: 6 de maio de 2022.

BAI, Y. et al. **Integrating Knowledge into End-to-End Speech Recognition from External Text-Only Data**. IEEE/ACM Transactions on Audio Speech and Language Processing. v. 29, p. 1340–1351, 2021.

CAVALCANTI, A. E. L. W.; FILHO, N. G. T. **Aplicativos de gestão e influenciadores financeiros nas redes sociais como mecanismos de propagação da educação financeira**. Revista Juris Poiesis, Rio de Janeiro. v. 24, n. 36, p. 01-20, 2021. Disponível em: <http://periodicos.estacio.br/index.php/jurispoiesis/article/viewFile/10267/47968147>. Acesso em: 16 de junho de 2022.

CIOFFI, R. et al. **Artificial intelligence and machine learning applications in smart production: Progress, trends, and directions**. Sustainability. v.12, n. 492. 2020. Disponível em: <https://www.mdpi.com/2071-1050/12/2/492>. Acesso em: 16 de junho de 2022.

CISCO. **Digitizing and packetizing voice**. Disponível em: <https://www.yumpu.com/en/document/read/50717104/lesson-22-digitizing-and-packetizing-voice>. Acesso em: 14 de abril de 2022.

CNC - DIVISÃO ECONÔMICA | RIO DE JANEIRO. **Endividamento a inadimplência do consumidor**. Disponível em: <https://portal-bucket.azureedge.net/wp-content/2021/11/Graficos_Peic_out_2021.pdf>. Acesso em: 18 de março de 2022.

COSTA, E. A. DE Q.; SOUZA, D. S.; AMARAL, I. DA S. DO. **GESTÃO DAS FINANÇAS PESSOAIS: UMA VIDA ECONOMICAMENTE CORRETA.** Ciências Humanas e Sociais, Aracaju, v. 6, n. n.3, p. 71–84, mar. 2021. Disponível em:

<https://periodicos.set.edu.br/cadernohumanas/article/view/7683>. Acesso em: 16 de junho de 2022.

CUADROS, C. D. R. **RECONHECIMENTO DE VOZ E DE LOCUTOR EM AMBIENTES RUIDOSOS: COMPARAÇÃO DAS TÉCNICAS MFCC E ZCPA**. Niterói, 2007. Disponível em: <http://aquarius.ime.eb.br/~apolin/papers/Carlos_UFF_2007.pdf>. Acesso em: 16 de junho de 2022.

DOSHI, K. **Audio Deep Learning Made Simple (Part 1): State-of-the-Art Techniques**. Disponível em: <https://towardsdatascience.com/audio-deep-learning-made-simple-part-1-state-of-the-art-techniques-da1d3dff2504>. Acesso em: 14 de abril de 2022.

Gomes, C. T. Pedro. **Análise de Sentimentos com Machine Learning**. Disponivel em: <https://www.datageeks.com.br/analise-de-sentimentos/>. Acesso em: 15 de dezembro de 2022.

IBM. **Introdução ao Aprendizado de Máquina**. Disponível em: <https://www.ibm.com/br-pt/analytics/machine-learning>. Acesso em: 9 de maio de 2022.

ISLAM, M. T.; ISLAM, B.; NIRJON, S. **SoundSifter: Mitigating overhearing of continuous listening devices**. MobiSys'17, Niagara Falls, NY, USA, p. 29-41, 2017.

Disponível em: < https://dl.acm.org/doi/10.1145/3081333.3081338>. Acesso em: 16 de junho de 2022.

LAPA, A. E. et al. **APLICATIVOS DE GESTÃO E INFLUENCIADORES FINANCEIROS NAS REDES SOCIAIS COMO MECANISMOS DE PROPAGAÇÃO DA EDUCAÇÃO FINANCEIRA MANAGEMENT APPLICATIONS AND FINANCIAL INFLUENCERS ON SOCIAL NETWORKS AS MECHANISMS FOR SPREADING FINANCIAL EDUCATION**. Revista Juris Poiesis, Rio de Janeiro. v. 24, n. 36, p. 01-20, 2021. Disponível em: <http://periodicos.estacio.br/index.php/jurispoiesis/article/viewFile/10267/47968147>. Acesso em: 16 de junho de 2022.

MOAR, J.; ESCHERICH, M. **HEY SIRI, HOW WILL YOU MAKE MONEY?** Disponível em: <https://www.juniperresearch.com/whitepapers/hey-siri-how-will-you-make-money>. Acesso em: 15 abr. 2022.

Nageshkar, P. **Multi-label Text Classification using Transformers(BERT)**. Disponivel em: <https://medium.com/analytics-vidhya/multi-label-text-classification-using-transformers-bert-93460838e62b>. Acesso em: 15 de dezembro de 2022.

NSC, ESTÚDIO. **Fintechs cresceram 34% em 2020, democratizando os serviços financeiros no Brasil**. Disponível em: <https://www.nsctotal.com.br/noticias/fintechs-cresceram-34-em-2020-democratizando-os-servicos-financeiros-no-brasil>. Acesso em: 18 mar. 2022.

OH, S. et al. **Command recognition using binarized convolutional neural network with voice and radar sensors for human-vehicle interaction**. Sensors, v. 21, n. 11, 5 de junho de 2021. Disponível em: <https://www.mdpi.com/1424-8220/21/11/3906>. Acesso em: 16 de junho de 2022.

OLIVEIRA, F. R. DE. **Afinal, o que é machine learning? Confira o conceito e exemplos!** Disponível em: <https://conteudo.movidesk.com/o-que-e-machine-learning/> Acesso em: 16 de junho de 2022.

REDAÇÃO. **Machine Learning é o futuro e outros destaques do Google I/O**. 13 maio 2019. Disponível em: <https://proximonivel.embratel.com.br/machine-learning-e-o-futuro-e-outros-destaques-do-google-i-o-a-conferencia-para-desenvolvedores-do-google/>. Acesso em: 16 de junho de 2022.

Rodrigues, V. **Métricas de Avaliação: acurácia, precisão, recall… quais as diferenças?**. Disponível em: <https://vitorborbarodrigues.medium.com/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-acur%C3%A1cia-precis%C3%A3o-recall-quais-as-diferen%C3%A7as-c8f05e0a513c>. Acesso em: 15 de dezembro de 2022.

ROEVER, L. **Compreendendo os estudos de revisão sistemática**. Rev Soc Bras Clin Med, abr-jun, p. 127-130, 2017. Disponível em: <https://www.sbcm.org.br/ojs3/index.php/rsbcm/article/view/276/255> Acesso em: 16 de junho de 2022.

SHIMON, C. et al. **Artificial intelligence enabled preliminary diagnosis for COVID-19 from voice cues and questionnaires**. The Journal of the Acoustical Society of America, v. 149, n. 2, p. 1120–1124, fev. 2021.

SOUSA, W. A. DE. **MUDANÇA TECNOLÓGICA NOS MEIOS DE PAGAMENTO. SERIA POSSÍVEL FAZER UMA AVALIAÇÃO SOBRE A REDUÇÃO DOS CUSTOS DE TRANSAÇÃO?** Osasco, 2021. Disponível em: <https://repositorio.unifesp.br/handle/11600/61563> Acesso em: 16 de junho de 2022.

SOARES De Paiva E, Pereira F. **Capítulo 2 Deep Learning para Processamento de Linguagem Natural**. Disponivel em: <https://sol.sbc.org.br/livros/index.php/sbc/catalog/download/86/378/642-1?inline=1>. Acesso em: 15 de dezembro de 2022.

SOUZA, I. DE. **Fala que eu te escuto**. Disponível em: <https://rockcontent.com/br/blog/comando-de-voz/>. Acesso em: 15 jun. 2022.

SUN, S. et al. **Adversarial regularization for attention based end-to-end robust speech recognition**. IEEE/ACM Transactions on Audio Speech and Language Processing, v. 27, n. 11, p. 1826–1838, 1 nov. 2019.

TECNOLOGIA, P. **Tecnologia de voz: entenda as implicações dessa tendência!** Disponível em: <https://www.meupositivo.com.br/panoramapositivo/tecnologia-de-voz/>. Acesso em: 14 abr. 2022.

VANDERLEY, M. S.; SILVA, J. G. DOS S.; ALMEIDA, S. A. DE. **Educação financeira na infância e adolescência e seus reflexos na vida adulta: uma revisão de literatura**. Faculdade de ciências do Tocantins, v. 1, n. Ed. 20, p. 146–166, nov. 2020.

ZHANG, G. et al. **DolphinAttack: Inaudible voice commands**. Proceedings of the ACM Conference on Computer and Communications Security. **Anais**...Association for Computing Machinery, 30 out. 2017.