INSTITUTO FEDERAL DE SANTA CATARINA

JOÃO LEONARDO MARTINS

APLICAÇÃO WEB PARA PROCESSAMENTO DE LINGUAGEM NATURAL EM PORTUGUÊS UTILIZANDO A TÉCNICA BERT

São José - SC

dezembro/2022

APLICAÇÃO WEB PARA PROCESSAMENTO DE LINGUAGEM NATURAL EM PORTUGUÊS UTILIZANDO A TÉCNICA BERT

Projeto de Trabalho de conclusão de curso apresentado à Coordenadoria do Curso de Engenharia de Telecomunicações do campus São José do Instituto Federal de Santa Catarina.

Orientador: Prof. Mário de Noronha Neto, Dr.

São José - SC

dezembro/2022

RESUMO

O emprego de técnicas de Inteligência Artificial (IA) tem permitido avanços muito relevantes em nosso cotidiano, nos mais diversos setores: desde assistentes virtuais, passando por anúncios e *feeds* personalizados em redes sociais, até chegar nos atendentes virtuais. Muito disso passa pela compreensão profunda das intenções de buscas e interações dos usuários, analisado atentamento pelos algoritmos utilizados por empresas, graças ao uso de técnicas de Processamento de Linguagem Natural (PLN). Neste trabalho busca-se entender os conceitos estudados nos principais ramos de IA relacionados ao PLN, com enfoque no algoritmo BERT (*Bidirecional Encoder Representations from Transformers*), da empresa Google, que teve seu código fonte aberto e é amplamente utilizado no buscador da mesma. De forma a observarmos o uso do BERT na prática, por fim, será desenvolvida uma API para que o modelo possa ser treinado (como é chamada a parte de refinamento do algoritmo) e assim seja capaz de classificar o sentimento por trás de frases em português, as identificando como negativas, neutras ou positivas.

Processamento de Linguagem Natural. BERT. API. Aplicação Web.

SUMÁRIO

1 INTRODUÇÃO 4
1.1 Objetivo Geral 5
1.2 Objetivos Específicos 5

2 FUNDAMENTAÇÃO TEÓRICA 6
2.1 Fundamentos gerais de IA 6
2.2 Aprendizado de Máquina 7
2.2.1 Aprendizado Profundo 8
2.2.1.1 Redes Neurais 8
2.2.1.2 Redes Neurais Convolucionais 11
2.2.1.3 Redes Neurais Recorrentes 11
2.2.1.4 *Long Short-Term Memory* 12
2.3 Processamento de Linguagem Natural 13
2.4 *Transformers* 15
2.4.1 *Bidirectional Encoder Representations from Transformers* 16

3 PROPOSTA 18

REFERÊNCIAS 19

1 INTRODUÇÃO

Recentemente pudemos acompanhar avanços importantes em Inteligência Artificial, cada vez mais presente no nosso cotidiano, seja em algoritmos de busca, detecção de doenças ou em atendimentos de empresas, através de *chatbots* (ALBAYRAK; ÖZDEMIR; ZEYDAN, 2018). Com a situação econômica de diversas empresas prejudicada pela pandemia, surge a necessidade de investimentos pontuais que impliquem num maior retorno com a menor contrapartida possível, é onde entra a IA, podendo ser aplicada em publicidades mais eficientes, de forma a impactar um público mais específico e com grande probabilidade de retorno (JAREK; MAZUREK, 2019).

Sempre quando há um avanço tecnológico que possibilite que uma tarefa passe a ser executada por uma máquina, há um receio geral de que pessoas perderão seus empregos. Porém, o que é possível observar é que, num primeiro momento, os frutos da IA são utilizados como auxiliares, como por exemplo, em um pré-atendimento, seja por web chat ou através da telefonia convencional. Isso traduz em produtividade que, por sua vez, possibilita a implantação de uma estratégia em voga: o conceito de *Omnichannel*, que é, basicamente, a convergência de todos os canais utilizados por uma empresa (SCHROTENBOER, 2019).

Quando utilizamos os buscadores modernos, fatalmente nos deparamos com IA. E com o que há de mais moderno, pode-se dizer. Por exemplo, ao acessar os mecanismos de buscas do Google, estamos em contato com um algoritmo chamado BERT, que é o que há de mais atual em Processamento de Linguagem Natural e que permite uma busca muito mais refinada, com resultados muito mais próximos ao que o usuário realmente se refere (daí o conceito de Processamento de Linguagem Natural). O autopreenchimento em buscas é mais um recurso possível graças ao emprego do BERT (DEVLIN et al., 2018).

Outro conceito interessante que observamos nos buscadores, porém não apenas nos mesmos, é o de Visão Computacional, que é o ramo de IA que estuda o processamento de imagens pelos computadores. Com isso, é possível realizarmos uma busca a partir de uma figura, ou então garantir a entrada de pessoas (reconhecimento facial), reconhecimento de padrões de modo a evitar acidentes, leitura de placas de forma automatizada, entre outras aplicações (MARTINEZ; AL-HUSSEIN; AHMAD, 2019).

Sobre PLN, como mencionado, houve um avanço bem interessante desde o início da pandemia de Covid-19, em que muitas empresas buscaram formas de automatizar o seu atendimento, fomentando o setor (ZAGO, 2021). Além do emprego em atendimento ao cliente e buscas, como já citado, o PLN pode ser aplicado em tradução de idiomas, filtro de SPAM, reconhecimento de fala, aplicativos assistentes pessoais, entre outros. Tudo isso teve início na década de 1940, quando surge uma maior necessidade, com os aprendizados

da Segunda Guerra Mundial, de máquinas que pudessem traduzir de uma língua para outra rapidamente.

Ainda que o conceito de PLN esteja se popularizando e cada vez mais presente em nossas vidas, há muitas possibilidades de aperfeiçoamento e implementação dessa tecnologia, que cada vez mais vem atuando como uma interface entre máquinas e humanos. Assim sendo, gradativamente iremos observar avanços nessa área, seja no aperfeiçoamento do entendimento das solicitações de usuários em buscadores, por exemplo, como também numa maior integração com outros segmentos de IA, além de outras áreas, como *Internet of Things*.

1.1 Objetivo Geral

Este trabalho objetiva o entendimento básico dos modelos modernos para PLN, assim como suas interações com outros campos de IA e, principalmente, suas aplicações e impactos na sociedade.

Será abordada com maior detalhamento a técnica BERT e, como resultado, espera-se obter uma ferramenta, baseada nesse modelo, que nos dê um embasamento para determinar se as pessoas estão gostando de determinada experiência (que pode ser um serviço ou um produto). Tudo isso com base num treinamento prévio e uma base de dados de avaliações dos mesmos.

1.2 Objetivos Específicos

Alguns pontos foram pensados como meta para este trabalho, de modo a atingir o entendimento buscado de forma geral:

1. Obtenção de uma base de dados para estudo de alguma aplicação em específico
2. Claro entendimento da fase de treinamento e reprodutibilidade para outras bases de dados
3. Desenvolvimento de uma API

2 FUNDAMENTAÇÃO TEÓRICA

Neste Capítulo 2 serão abordados, com maiores detalhes, os assuntos pertinentes ao desenvolvimento deste trabalho. Foram realizadas revisões bibliográficas para conceitos como Inteligência Artificial, de uma forma geral, bem como *Natural Language Processing* e *Transformers*.

2.1 Fundamentos gerais de IA

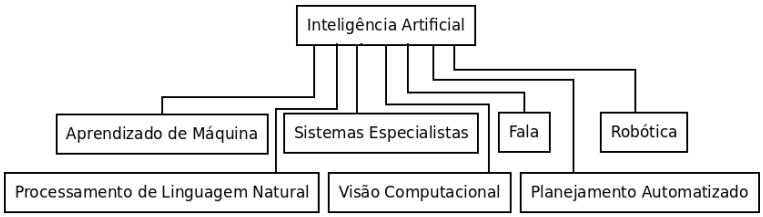
IA é uma parte da Ciência da Computação que, principalmente, trabalha com a questão da automação do comportamento inteligente, seja do domínio humano, animal ou vegetal, conforme Chowdhary (2020). Nesse contexto, comportamento inteligente é definido pelo autor como uma combinação de percepção, análise e reação. Na prática, isso tudo implica em ricas experiências para os usuários de assistentes virtuais (como a Siri e a Alexa), carros autônomos ou semiautomatizados, mecanismos de buscas e tantos outros que diariamente temos contato.

É válido afirmar que essas diversas facilidades são possíveis graças aos recursos de IA, que permitem sistemas aprenderem através de experiências e memorizá-las, sendo capazes de tomarem decisões com base num histórico e até mesmo preverem situações. Além do mais, apresentam traços de inteligência humana, através de interação por linguagem natural. São resultados da combinação de avanços de hardware e software, por meio de evolução do poder de processamento e algoritmos, respectivamente (XAVIER, 2020).

De acordo com Villanueva e Salenga (2018), há sete áreas de estudo em IA, conforme a Figura 1. Cada subconjunto, combinados ou não, permitem o desenvolvimento de facilidades citadas nesta seção e inúmeras outras. Exemplificando o produto de algumas das principais áreas de IA, podemos citar para o domínio de Aprendizado de Máquina os alertas de tráfego, muito comuns nos aplicativos de mapas. Para o âmbito de de Processamento de Linguagem Natural, uma forma de emprego é para análise de sentimento, como veremos mais adiante. Já para Sistemas Especialistas, através de detectores de vírus.

Além dos campos de IA mencionados, há também a possibilidade de divisão sob outra perspectiva, segundo a abordagem de Searle (1980), que são os conceitos de IA Geral e IA Estreita, ou IA Forte e IA Fraca, respectivamente. A primeira traz a ideia de uma IA capaz de raciocinar de fato, atuando como uma mente propriamente dita. Já a Inteligência Artificial Fraca é caracterizada como uma ferramenta para resolver problemas pontuais. Naturalmente, a segunda forma está muito mais desenvolvida, sendo possível observá-la através das aplicações citadas anteriormente, enquanto a primeira, por se tratar de um

Figura 1 – Áreas de estudo de IA



Fonte: O próprio autor

ramo que visa dar consciência às máquinas, ainda é bastante prematura e incerta.

2.2 Aprendizado de Máquina

Machine Learning em inglês é o segmento de IA que busca-se programar computadores, não necessariamente de uma forma explícita, para que possam aprender como seres humanos, mas como não são capazes de raciocinar, é correto afirmar que o aprendizado vem de treinamentos. Portanto, esse conhecimento não vem apenas do armazenamento de (muitas vezes) grandes quantidades de dados, mas sim da extração de padrões de variadas formas de fontes de dados, como, por exemplo, vídeos, interações de redes sociais, falas de seres humanos, dados climáticos, entre outros (KAPLAN, 2016).

De acordo com Wakefield (2019) e Bernard (2021), há quatro paradigmas de aprendizagem principais, que basicamente diferem na forma em que os dados são fornecidos ao computador e no tipo de problema que podem resolver. São eles:

1. **Aprendizagem supervisionada**: esse modelo é o mais aplicado atualmente. Em suma, a máquina aprende através de exemplos rotulados, em que as entradas e saídas são conhecidas. Com isso, o algoritmo deve encontrar uma forma de como chegar nos resultados de acordo com os dados fornecidos. Seu nome se dá pelo fato de que, enquanto o código não obtém um nível satisfatório de acerto em suas tarefas, o operador atua de modo a refiná-lo. Pode ser utilizado para realizar **classificações**, cujo objetivo é definir qual categoria correspondente às novas ocorrências (com base num conjunto de dados) ou então para **regressões**, que buscam traduzir em números alguma questão, portanto, seu trabalho é prever um valor numérico;

2. **Aprendizagem não supervisionada**: apesar de não ser tão utilizado quanto a supervisionada, a aprendizagem não supervisionada é considerada pelos autores como chave para o desenvolvimento de sistemas artificialmente inteligentes. Este modelo é utilizado em situações mais específicas e funciona, basicamente, analisando conjuntos de dados e identificando padrões, buscando relações entre eles de modo a catalogá-los. Isso é chamado de **agrupamento** e é utilizado para separar indivíduos

por seus gostos para utilização em campanhas de marketing, por exemplo. Pode ser utilizada também para **redução de dimensionalidade**, que é uma técnica de aprendizagem não supervisionada empregada para restrição de variáveis dentro de bases de dados, preservando o que realmente importa. Pode ser observada em sistemas de recomendações;

3. **Treinamento por reforço**: trata-se de um modelo de aprendizagem consideravelmente diferente dos anteriores, tendo em vista que a aquisição dos conjuntos de dados não ocorre manualmente através de entradas convencionais, mas sim através de interações com outros sistemas, chamados de **ambiente** neste contexto. Simplificando, pode-se dizer que a máquina aprende através de tentativa e erro, capacitando-se através de experiências passadas e se adaptando através dos resultados obtidos. Uma aplicação deste modelo pode ser observada em alguns data centers da empresa Google, em que o sistema de arrefecimento é controlado por algoritmos treinados por reforço, e os resultados são positivos, tendo em vista que o consumo energético foi otimizado;

4. **Treinamento por transferência**: é uma forma de aprendizagem que busca associar um conhecimento já adquirido em determinada tarefa em outra, facilitando a absorção do conteúdo. É similar ao que os seres humanos fazem o tempo todo: nós reutilizamos nossos saberes de modo a acelerar novos afazeres. Em IA, treinamento por transferência é utilizado no ramo de Visão Computacional - através de comparações de imagens e no campo de PLN, sintetizando textos, por exemplo.

2.2.1 Aprendizado Profundo

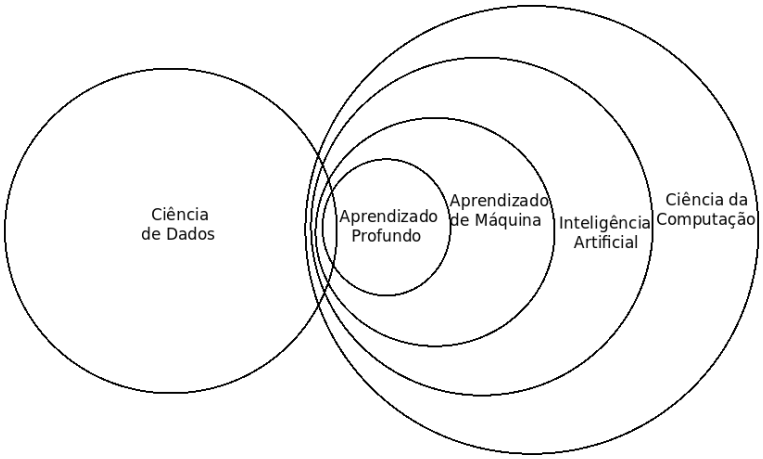
Em inglês *Deep Learning*, pode ser definido como uma técnica para implementação de *Machine Learning*, técnica essa que permite a utilização em larga escala de ML através de seus algoritmos, que dividem tarefas de formas simples. Isso tudo é possível graças às redes neurais, que formam diversas camadas de processamento e possibilitam o aprendizado através de um grande volume de dados. São aplicações de DL os campos de Visão Computacional e PLN. Um exemplo clássico de uma aplicação prática proporcionada por DL foram vídeos de gatos no YouTube sendo classificados e sugeridos aos usuários de forma automática (ONGSULEE, 2017).

Conforme a Figura 2, DL é uma das formas possíveis de execução de ML, que por sua vez é uma abordagem para chegar à *Artificial Intelligence*.

2.2.1.1 Redes Neurais

Conforme adiantado na seção 2.2.1, as redes neurais artificiais possibilitam a implementação de sistemas artificialmente inteligentes, e isso é alcançado através redes

Figura 2 – Relação de Aprendizado Profundo



Fonte: O próprio autor

neurais, que são inspiradas no cérebro humano, ou melhor, no modo em que é composto, através de neurônios, e organizado, de movo a formar uma rede. De forma semelhante, as neurônios artificiais se comunicam, permitindo o processamento de um volume grande de informação (BERNARD, 2021).

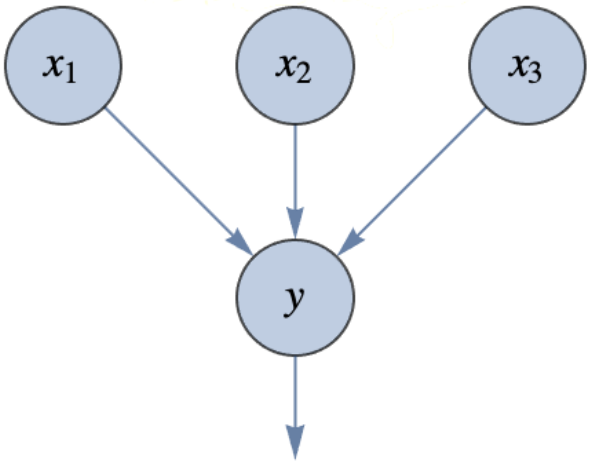
A Figura 3 representa, sinteticamente, a forma mais básica das redes neurais, que são chamadas de **unidades**, ou unidades de processamento. Nela podemos observar, denominados por x, os sinais de entrada aplicados à célula artificial, que, após cálculos envolvendo uma soma ponderada, de acordo com a influência dos sinais, é apresentada uma saída y (BERNARD, 2021). Outra nomenclatura comum é *Perceptron*, de acordo com o desenvolvedor Frank Rosenblatt, o cientista responsável e que foi inspirado pelos trabalhos anteriores de Warren McCulloch e Walter Pitts. A invenção ocorreu nas décadas de 1950 e 1960 (ACADEMY, 2019).

De forma matemática, um neurônio artificial pode ser representado pela equação Equação 2.1, em que a saída y é o resultado da função não linear f, dada pela multiplicação de parâmetros w - que podem ser aprendidos - e são chamados de **pesos**, multiplicados pelas entradas. É somado a isso tudo o parâmetro b, chamado de **bias**, ou viés, que pode restringir ou aumentar os pesos.

$$y = f(w1x1 + w2x2 + w3x3 + b) \tag{2.1}$$

A partir do agrupamento de neurônios artificiais em rede é possível chegarmos às **redes neurais**, que, em suma, se dá pela conexão de muitos desses neurônios. Essa ligação entre eles não é feita de forma randômica e utiliza uma arquitetura clássica, que pode ser definida com a essência de um modelo de DL, conforme Bengio (2016), que é a ***feed-forward neural network*** ou rede neural de alimentação para frente, caracterizada por ser dividida em níveis, com neurônios artificiais ligados às camadas imediatamente

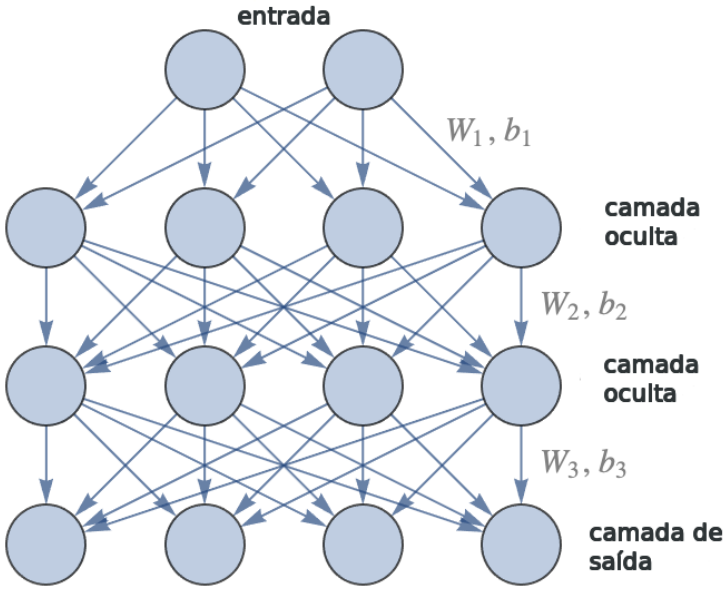
Figura 3 – Representação de um Neurônio Artificial



Fonte: Adaptada de Bernard (2021)

anterior e superiores (e não às próprias ou adjacentes), possuir N camadas intermediárias . Uma ilustração pode ser observada na Figura 5.

Figura 4 – Representação clássica de uma Rede Neural



Fonte: Adaptada de Bernard (2021)

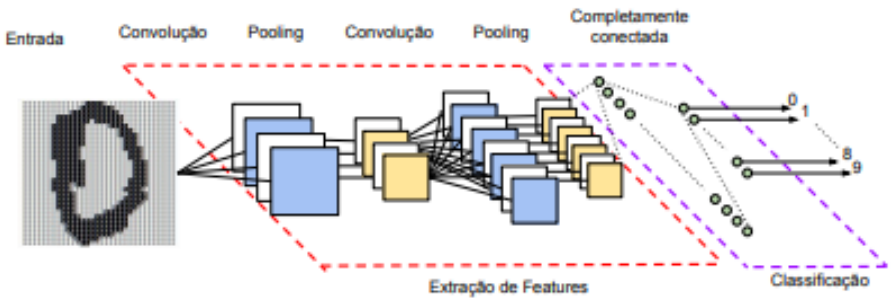
Com a interligação de milhões de neurônios artificiais, inúmeras camadas ocultas de redes neurais podem ser formadas, dando origem às **redes neurais profundas**. Daí o nome *Deep Learning*, em que a palavra *deep*, ou profundo, se refere ao fato da utilização de redes neurais com o mesmo nome: *deep neural networks* (BERNARD, 2021).

2.2.1.2 Redes Neurais Convolucionais

É uma das arquiteturas de redes neurais profundas mais clássicas. Também conhecida por *Classical Neural Networks*, é uma variação da rede neural de alimentação para frente, vista na seção 2.2.1.1. Difere no sentido de que, como o nome sugere, utiliza a operação matemática **convolução**. Para tanto, ao menos uma camada de suas redes utilizam essa operação no lugar de multiplicação. É especialmente utilizada no campo de Visão Computacional, podendo ser empregada em agrupamento de imagens, busca por fotos ou reconhecimento de objetos, por exemplo. (ZEILER; FERGUS, 2014).

A Figura 5 representa as etapas que compõem uma CNN, cabendo a cada uma dessas camadas desempenharem uma função na propagação do sinal. São elas: **convolução**, responsável pela definição dos atributos de entrada através de filtragens realizadas por neurônios, que resultam em um conjunto de pixels encaminhados ao nível seguinte; ***pooling*** ou agrupamento, que simplifica o que foi obtido anteriormente, reduzindo o número de amostras; já a camada **completamente conectada** tem como objetivo ligar os neurônios de *pooling* à saída, classificando o resultado de acordo com as respostas dos filtros das camadas anteriores.

Figura 5 – Arquitetura de uma Rede Neural Convolucional



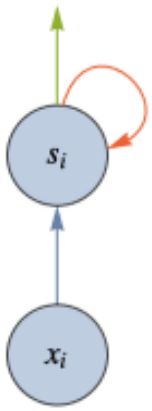
Fonte: Vargas, Paes e Vasconcelos (2016)

2.2.1.3 Redes Neurais Recorrentes

As *Recurrent Neural Networks*, por sua vez, são amplamente em PLN, processando informações sequenciais como, por exemplo, texto e áudio. Da mesma forma que as CNN, as RNN também aprendem através de treinamento, porém com um diferencial de utilizar informações passadas para tomar decisões, ou seja, possui memória (BENGIO, 2016).

Conforme (BERNARD, 2021), uma Rede Neural Recorrente pode ser simplificada como um módulo processando dados sequenciais enquanto mantém e atualiza um estado, estado esse que, como citado, poderá ser reutilizado. A Figura 6 ilustra essa forma mais básica de Rede Recorrente, em que x e s, que são, respectivamente, vetores numéricos da informação sequencial de entrada e do estado, respectivamente.

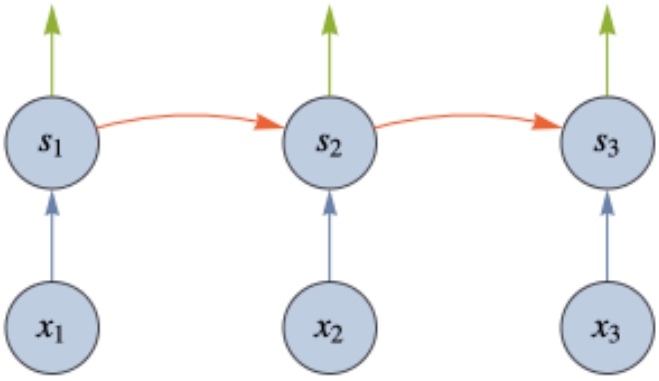
Figura 6 – Camada de recorrência



Fonte: Bernard (2021)

Já a Figura 7 exibe os desdobramentos para uma sequência exemplificada com comprimento três, deixando claro a característica de influência do estado anterior no cômputo do estado atual.

Figura 7 – Rede de recorrência



Fonte: Bernard (2021)

2.2.1.4 *Long Short-Term Memory*

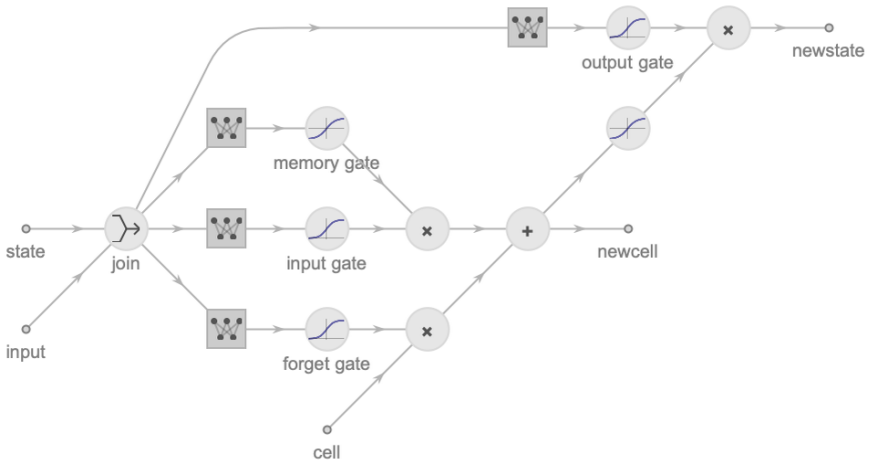
De acordo com Graves (2012), as LSTM vieram para superar as limitações das RNN, contornando as perdas de influência que uma determinada entrada sofre em grandes sequências de dados. Simplificando, são capazes de aprender conexões de longo prazo. Tudo isso é possível através de **portas**, ou *gates*, em inglês, que possuem a capacidade de definir o que será armazenado ou eliminado das memórias internas.

Na Figura 8 temos detalhes da estrutura da rede LSTM, formadas pelas seguintes portas: ***forget gate***, ***input gate***, ***memory gate*** e ***output gate***, pelos vetores numéricos

de entrada e saída, ***input*** e ***output*** e pelos vetores de estado ***state*** e ***cell state***, sendo esse último utilizado para levar informação a etapas posteriores. Seu funcionamento ocorre da seguinte forma:

- A porta *forget*, multiplicando *cell*, é capaz de definir o que deve ser levado em conta e o que deve ser esquecido da memória de longo prazo;
- *Input gate*, é multiplicada pela *memory gate*, e posteriormente somada à porta de estado *cell*. Ou seja, *memory gate* calcula o que deve ser adicionado à *cell state* e *input gate* decide se de fato essa adição deve ocorrer;
- *output gate* define quais valores de *cell state* devem ser levados ao próximo estado.

Figura 8 – Núcleo da rede da arquitetura LSTM



Fonte: Bernard (2021)

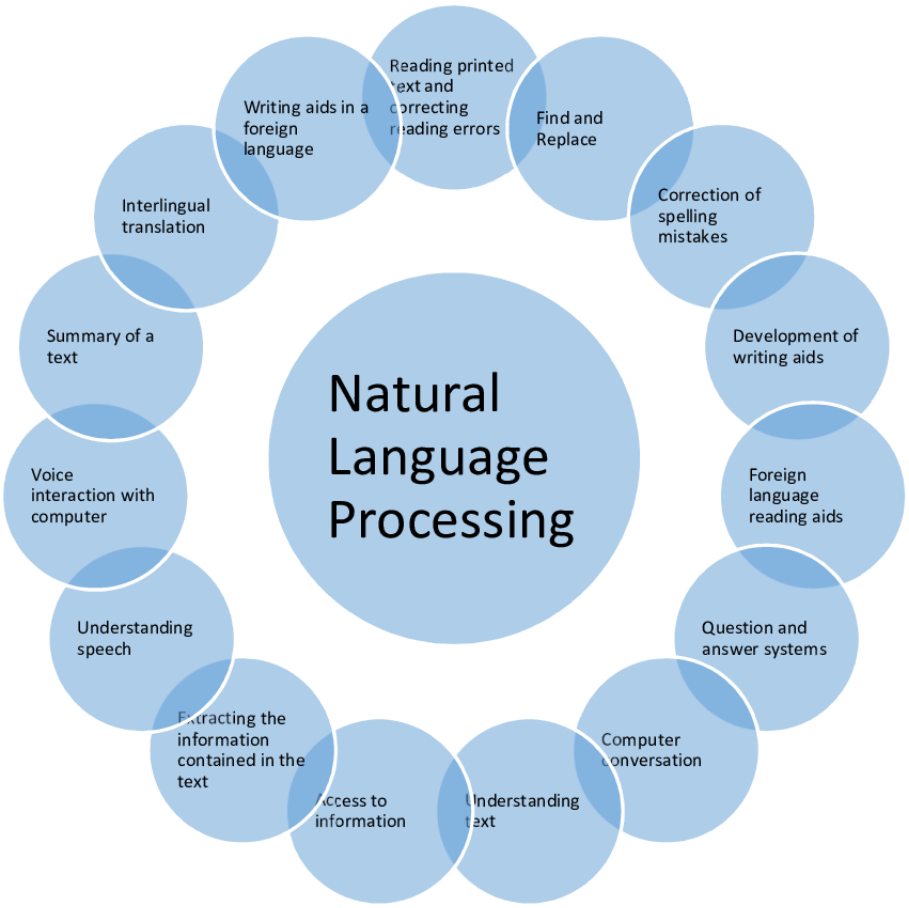
2.3 Processamento de Linguagem Natural

Em inglês *Natural Language Processing*, é o ramo de IA que trata de capacitar as máquinas para compreenderem o processo natural de comunicação dos seres humanos, não apenas realizando traduções livres mas sim observando suas nuances nos mais variados contextos (EISENSTEIN, 2019).

Aplicações como *Text to Speech*, Análise de Sentimentos, Assistentes Virtuais entre outras diversas, conforme a Figura 9, são viáveis através do PLN.

De acordo com Liddy (2001), há quatro tipos de abordagem, que exigem, naturalmente, diferentes níveis de processamento:

Figura 9 – Tópicos de PLN



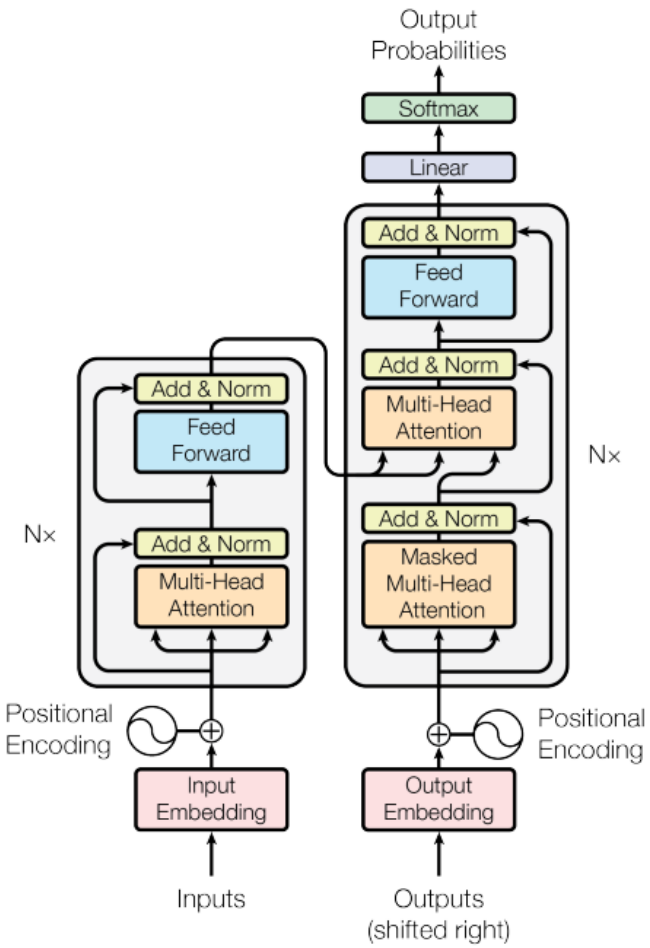
Fonte: Sutçu e Aytekin (2019)

- **Simbólica**, que é baseada em regras de linguagem bem limitadas, ou seja, que não tenham ambiguidade. Algoritmos são construídos de modo a realizar classificações simples de palavras, uma a uma;
- A **estatística** baseia-se em sistemas supervisionados, conforme seção 2.2, e aplica modelos matemáticos, como, por exemplo, cadeias de Markov. Diferente da simbólica, a abordagem estatística é capaz de realizar deduções de interpretação, dispensando assim regras linguísticas;
- Já a abordagem **conexionista** é similar à estatística, com o diferencial de se valer da junção de teorias de conhecimento com aprendizado estatístico, dando maior liberdade para dedução, transformação e manipulação de textos;
- Por último, a **híbrida**, como o nome sugere, combina todas as anteriores, dando um grau ainda maior de flexibilidade para resolução de problemas de PLN.

2.4 *Transformers*

Transformer é uma arquitetura de rede neural baseada no mecanismo de **atenção**, que é, basicamente, um processo similar ao que a mente humana faz: em redes neurais profundas, atenção ignora partes irrelevantes, enquanto se concentra em outras. (HAN et al., 2021) Com isso, *Transformer* surge como uma excelente alternativa às redes neurais recorrentes estudadas na subseção 2.2.1.3, que possuem limitações, como, por exemplo, falta de efetividade em modelo cuja entrada e saída são sequencialmente distantes. Como a arquitetura *Transformer* utiliza uma estrutura do tipo *encoder-decoder*, há a possibilidade de uma paralelização em seu treinamento, se tornando muito mais efetivo para as tarefas de Processamento de Linguagem Natural (VASWANI et al., 2017). A Figura 10 exibe o modelo da arquitetura *Transformer*.

Figura 10 – Arquitetura modelo - *Transformer*



Fonte: Vaswani et al. (2017)

De acordo com Vaswani et al. (2017), além do mecanismo de atenção, *Transformer* se destaca através do uso de ***positional encodings***, que permitem ao modelo conhecer as posições das palavras sem a necessidade do uso de recorrência ou convolução, aumentando

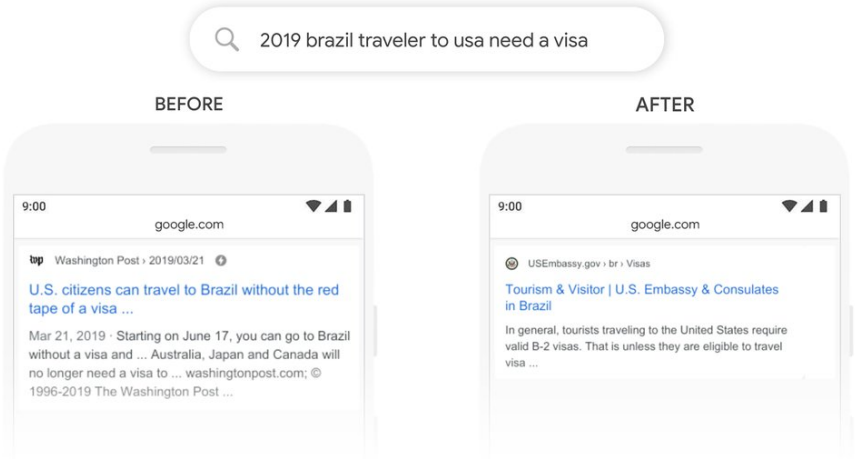
seu desempenho.

2.4.1 *Bidirectional Encoder Representations from Transformers*

O BERT é um dos mecanismos que a empresa Google utiliza no seu sistema de busca e é especialmente utilizado para compreender de uma forma eficaz o real sentido do que está sendo procurado, ou seja, o BERT é o algoritmo de *Deep Learning* utilizado pela multinacional (NAYAK, 2019). Como o BERT utiliza *Transformers*, há um entendimento maior entre as palavras submetidas ao modelo, possibilitando uma maior inteligência do contexto a que se referem. (DEVLIN et al., 2018)

Um exemplo prático do emprego do BERT pode ser observado na Figura 11, em que o buscador pôde captar o que realmente o usuário buscava e retornar a melhor resposta. Neste caso, a grande diferença se dá pela importância dada à palavra *to* e sua relação com as demais, compreendendo que o usuário não deseja viajar para o Brasil e sim é um brasileiro que deseja ir aos Estados Unidos.

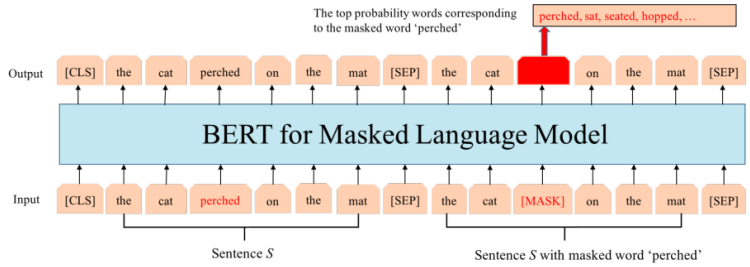
Figura 11 – NSP



Fonte: Nayak (2019)

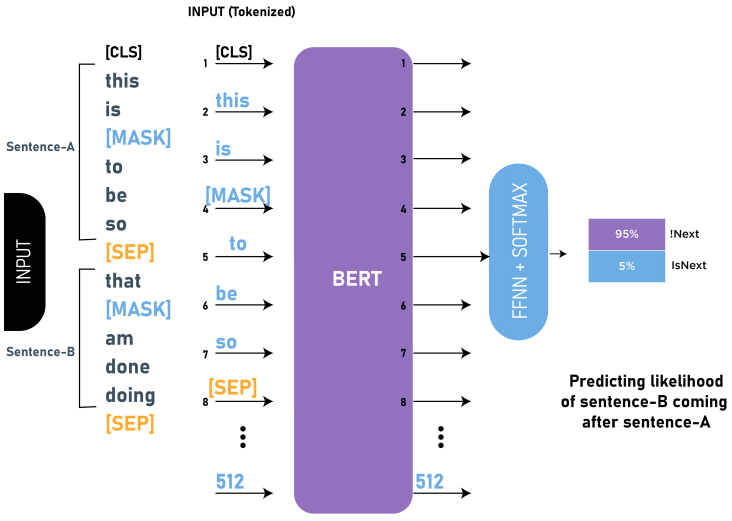
BERT é pré-treinado de por duas tarefas de Processamento de Linguagem Natural: *Masked-Language Modeling* e *Next Sentence Prediction*, sendo a primeira, como o nome sugere, um treinamento baseado na ocultação de uma palavra, em uma determinada frase, fazendo com que o software preveja o que foi mascarado dentro do contexto. Já a segunda, o modelo deve prever se dadas duas palavras têm relação ou se são simplesmente aleatórias (DEVLIN et al., 2018). A Figura 13 e a Figura 12, ilustram, respectivamente, o esquema de MLM e NSP.

Figura 12 – Modelo de linguagem mascarada



Fonte: Asthana (2021)

Figura 13 – Modelo de predição de próxima sentença



Fonte: Qiang et al. (2019)

3 PROPOSTA

A ideia do projeto é desenvolver uma ferramenta que auxilie pessoas a saberem se determinado assunto, que pode ser um produto ou serviço, por exemplo, estão agradando o público. Isso será feito através de um sistema em que o BERT será treinado através de um conjunto de dados. Para que funcione, esse *dataset* deve relacionar classificações ou notas a comentários, similar ao que é encontrado em avaliações de lojas de aplicativos. Com essas informações, o usuário terá condição de treinar o modelo para que o mesmo tenha a capacidade de, com base em seu treinamento prévio, deduzir se determinada frase é positiva, negativa ou neutra.

O objetivo é a implantação deste sistema como REST API, em que o usuário terá liberdade de incorporar o analisador em outros sistemas ou utilizar até mesmo via web, possibilitando, em tempo real, analisarmos a impressão do público sobre determinada situação. Para tanto, será utilizada linguagem Python e seus *frameworks* e bibliotecas:

- *Pytorch* - estrutura de Aprendizado Profundo
- *Transformers* - biblioteca com modelos pré-treinados
- *FastAPI* - *framework* web para desenvolvimento da API

Tabela 1 – Cronograma das atividades previstas

Etapa	Meses						
	Jan	Fev	Mar	Abr	Mai	Jun	Jul
Estudo sobre as ferramentas envolvidas	√	√	√	√			
Coleta de dados					√		
Desenvolvimento do *backend*						√	
Desenvolvimento da API							√

REFERÊNCIAS

ACADEMY, D. S. *Deep Learning Book*. 2019. Disponível em: <https://www.deeplearningbook.com.br/>. Acesso em: 25 de nov. de 2022. Citado na página 9.

ALBAYRAK, N.; ÖZDEMIR, A.; ZEYDAN, E. An overview of artificial intelligence based chatbots and an example chatbot application. In: IEEE. *2018 26th signal processing and communications applications conference (SIU)*. [S.l.], 2018. p. 1–4. Citado na página 4.

ASTHANA, P. *An application of sentiment analysis with transformer models on online news articles covering the Covid-19 pandemic*. [S.l.]: University of California, Los Angeles, 2021. Citado na página 17.

BENGIO, Y. *Deep Learning*. London, England: MIT Press, 2016. (Adaptive Computation and Machine Learning series). Citado 2 vezes nas páginas 9 e 11.

BERNARD, E. *Introduction To Machine Learning*. Champaign, IL: Wolfram Media, 2021. Citado 6 vezes nas páginas 7, 9, 10, 11, 12 e 13.

CHOWDHARY, K. R. *Fundamentals of artificial intelligence*. 1. ed. New Delhi, India: Springer, 2020. Citado na página 6.

DEVLIN, J. et al. BERT: Pre-training of deep bidirectional transformers for language understanding. out. 2018. Citado 2 vezes nas páginas 4 e 16.

EISENSTEIN, J. *Introduction to natural language processing*. London, England: MIT Press, 2019. (Adaptive Computation and Machine Learning series). Citado na página 13.

GRAVES, A. *Supervised sequence labelling with recurrent neural networks*. 2012. ed. Berlin, Germany: Springer, 2012. (Studies in computational intelligence). Citado na página 12.

HAN, K. et al. Transformer in transformer. In: RANZATO, M. et al. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2021. v. 34, p. 15908–15919. Disponível em: <https://proceedings.neurips.cc/paper/2021/file/854d9fca60b4bd07f9bb215d59ef5561-Paper.pdf>. Citado na página 15.

JAREK, K.; MAZUREK, G. Marketing and artificial intelligence. *Central European Business Review*, v. 8, n. 2, 2019. Citado na página 4.

KAPLAN, J. *Artificial intelligence*. New York, NY: Oxford University Press, 2016. (What Everyone Needs To Know (R)). Citado na página 7.

LIDDY, E. D. Natural language processing. 2001. Citado na página 13.

MARTINEZ, P.; AL-HUSSEIN, M.; AHMAD, R. A scientometric analysis and critical review of computer vision applications for construction. *Automation in Construction*, Elsevier, v. 107, p. 102947, 2019. Citado na página 4.

NAYAK, P. *Understanding searches better than ever before*. 2019. Disponível em: <https://www.blog.google/products/search/search-language-understanding-bert/>. Acesso em: 6 de dez. de 2022. Citado na página 16.

ONGSULEE, P. Artificial intelligence, machine learning and deep learning. In: *2017 15th International Conference on ICT and Knowledge Engineering (ICTKE)*. [S.l.: s.n.], 2017. p. 1–6. Citado na página 8.

QIANG, J. et al. A simple bert-based approach for lexical simplification. 2019. Citado na página 17.

SCHROTENBOER, D. The impact of artificial intelligence along the customer journey: a systematic literature review. University of Twente, 2019. Citado na página 4.

SEARLE, J. R. Minds, brains, and programs. *Behavioral and Brain Sciences*, Cambridge University Press, v. 3, n. 3, p. 417–424, 1980. Citado na página 6.

SUTçU, C. S.; AYTEKIN, C. An example of pragmatic analysis in natural language processing: Sentimental analysis of movie reviews. *Turk. Online J. Des. Art Commun.*, Turkish Online Journal of Design, Art and Communication (TOJDAC), CTC, n. 2019, p. 61–74, dez. 2019. Citado na página 14.

VARGAS, A. C. G.; PAES, A.; VASCONCELOS, C. N. Um estudo sobre redes neurais convolucionais e sua aplicação em detecção de pedestres. In: SN. *Proceedings of the xxix conference on graphics, patterns and images*. [S.l.], 2016. v. 1, n. 4. Citado na página 11.

VASWANI, A. et al. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017. Citado na página 15.

VILLANUEVA, M. B.; SALENGA, M. L. M. Bitter melon crop yield prediction using machine learning algorithm. *International Journal of Advanced Computer Science and Applications*, Science and Information (SAI) Organization Limited, v. 9, n. 3, 2018. Citado na página 6.

WAKEFIELD, K. *A guide to machine learning algorithms and their applications*. 2019. Disponível em: <https://www.sas.com/en$_g$b/*insights*/*articles*/*analytics*/*machine* − *learning* − *algorithms.html* > .*Acessoem* : 20*denov.de*2022.*Citadonapágina*7.

XAVIER, F. C. *Fundamentos, pilares e aplicações da Inteligência Artificial no setor público - MIT Technology Review*. 2020. Disponível em: <https://mittechreview.com.br/fundamentos-pilares-e-aplicacoes-da-inteligencia-artificial-no-setor-publico/>. Citado na página 6.

ZAGO, B. *Panorama de Atendimento online pré e pós-pandemia E tendências para 2021*. 2021. Disponível em: <https://www.ecommercebrasil.com.br/artigos/panorama-de-atendimento-online-pre-e-pos-pandemia-e-tendencias-para-2021/>. Acesso em: 24 de jul. de 2022. Citado na página 4.

ZEILER, M. D.; FERGUS, R. Visualizing and understanding convolutional networks. In: FLEET, D. et al. (Ed.). *Computer Vision – ECCV 2014*. Cham: Springer International Publishing, 2014. p. 818–833. ISBN 978-3-319-10590-1. Citado na página 11.