Instituto Superior de Tecnologias Avançadas
Licenciatura em Engenharia Informática

CleverSupport -
APLICAÇÃO DE LARGE LANGUAGE MODELS PARA IDENTIFICAR O SENTIMENTO EM E-MAILS DE TICKETING

Projeto Académico realizado por:

Jorge Fernando Teixeira Ramos - 2341

Coordenador: Professor Dr. Pedro Brandão

Orientadora: Professora Dra. Dulce Mourato

Lisboa, Ano Letivo 2022/2023

Instituto Superior de Tecnologias Avançadas
Licenciatura em Engenharia Informática

CleverSupport -
APLICAÇÃO DE LARGE LANGUAGE MODELS PARA IDENTIFICAR O SENTIMENTO EM E-MAILS DE TICKETING

Projeto Académico realizado por:

Jorge Fernando Teixeira Ramos - 2341

Coordenador: Professor Dr. Pedro Brandão

Orientadora: Professora Dra. Dulce Mourato

Lisboa, Ano Letivo 2022/2023

Agradecimentos

Gostaria de começar por expressar a minha mais profunda gratidão a todos aqueles que tornaram possível a realização deste projeto.

A minha gratidão à minha família: aos meus pais, irmã e à minha namorada Jamila, pelo apoio incondicional, compreensão e encorajamento ao longo de toda a minha esta jornada.

À Professora Dra. Dulce Mourato, minha orientadora, pela sua incansável orientação, apoio e paciência. O seu encorajamento e sabedoria foram pilares fundamentais na consecução deste trabalho.

Foi certamente um processo moroso e desafiante e que não seria possível sem o vosso apoio. A todos vocês, muito obrigado.

Resumo

Nos tempos recentes, os avanços em inteligência artificial, especialmente os Large Language Models (LLMs) como o ChatGPT, têm proporcionado uma transformação profunda na maneira como processamos e interpretamos a linguagem natural. Estas ferramentas modernas têm oferecido soluções mais ágeis e precisas para a análise de sentimentos em textos, o que é crucial em aplicações empresariais para entender feedback de clientes e otimizar o atendimento.

O primeiro pilar do projeto foi o desenvolvimento de uma ferramenta específica de análise. Esta aplicação foi construída para avaliar sentimentos presentes em e-mails. O objetivo era claro: auxiliar softwares de ticketing, na identificação e compreensão do sentimento embutido nas respostas dos clientes. Ao longo do desenvolvimento, foi dada uma atenção especial à simplicidade, tanto em termos de interface do usuário quanto na arquitetura de back-end. Assim, para materializar esta visão, optou-se pela framework ASP.NET WebForms, que proporcionou uma base sólida e confiável para o projeto. Complementando essa escolha, o Microsoft Azure com base de dados em SQL foi integrado para oferecer uma solução de armazenamento de dados segura e escalável.

O segundo pilar, igualmente crítico, estava focado na comparação e validação de diferentes modelos LLM na tarefa de análise de sentimentos. Cada comunicação, neste caso e-mails, foi avaliada de forma individual por três modelos: GPT-4, GPT-3.5 e GPT-3. Para assegurar uma análise rigorosa e objetiva, houve um esforço manual considerável. Cada e-mail foi classificado em categorias de sentimento, como Positivo, Neutro e Negativo. Uma vez feita essa classificação humana, ela serviu como uma espécie de padrão ouro para ser comparada com as avaliações geradas pelos modelos LLM.

Os resultados foram intrigantes. O GPT-3.5 emergiu como o modelo mais preciso, seguido de perto pelo GPT-4. No entanto, o GPT-3, contrariamente às expectativas, mostrou-se significativamente menos eficaz. Estas descobertas não só reforçaram a importância de uma seleção criteriosa de modelo, mas também sinalizaram que, por mais promissores que sejam, os LLMs ainda têm um caminho de aperfeiçoamento pela frente.

Palavras-chave: Inteligência Artificial, Análise de sentimentos, E-mails, Modelos LLM, ASP.NET WebForms.

Abstract

In recent times, advancements in artificial intelligence, especially Large Language Models (LLMs) like ChatGPT, have brought about a profound transformation in the way we process and interpret natural language. These modern tools offer quicker and more accurate solutions for sentiment analysis in texts, which is crucial in business applications to understand customer feedback and optimize service.

The first pillar of the project was the development of a specific analysis tool. This application was built to assess sentiments present in emails. The goal was clear: to assist ticketing software in identifying and understanding the sentiment embedded in customer responses. Throughout the development, special attention was given to simplicity, both in terms of user interface and back-end architecture. To realize this vision, the ASP.NET WebForms framework was chosen, providing a solid and reliable foundation for the project. Complementing this choice, Microsoft Azure with SQL-based data storage was integrated to provide a secure and scalable data storage solution.

The second pillar, equally critical, focused on comparing and validating different LLM models for sentiment analysis. Each communication, in this case, emails, was individually assessed by three models: GPT-4, GPT-3.5, and GPT-3. To ensure thorough and objective analysis, there was considerable manual effort. Each email was classified into sentiment categories such as Positive, Neutral, and Negative. Once this human classification was done, it served as a kind of gold standard to be compared with evaluations generated by the LLM models.

The results were intriguing. GPT-3.5 emerged as the most accurate model, closely followed by GPT-4. However, GPT-3, contrary to expectations, proved significantly less effective. These findings not only emphasized the importance of careful model selection but also signaled that, as promising as they might be, LLMs still have room for improvement ahead.

Keywords: Artificial Intelligence, Sentiment Analysis, Emails, LLM Models, ASP.NET WebForms.

Índice

Índice de Figuras

Índice de Tabelas

Introdução

Nos últimos tempos, observou-se a emergência dos Large Language Models (LLMs) a revolucionar o campo do processamento de linguagem natural, particularmente na análise de sentimentos de textos variados, como e-mails e feedbacks de clientes em softwares de ticketing. Resultado deste avanço, optou-se por este tema para o projeto acadêmico devido à sua relevância atual e ao potencial de impacto prático na melhoria da interação entre empresas e clientes.

Verificou-se que a eficiência dos LLMs em detetar e interpretar sentimentos expressos em textos, de forma automatizada e precisa, contrastava significativamente com as abordagens manuais, muitas vezes lentas e propensas a erros. Teoricamente, a capacidade dos LLMs de compreender nuances da linguagem humana oferecia uma rica base de análise para pesquisadores e profissionais. Quanto ao artefacto digital deste projeto, teve em vista desenvolver uma ferramenta baseada em LLMs que facilitasse e otimizasse a análise de sentimentos em e-mails de clientes, permitindo às empresas identificar e responder rapidamente a preocupações, elogios ou críticas.

Escolheu-se este projeto não apenas pela sua relevância teórica, mas também pelo seu potencial prático. As empresas estando constantemente em busca de melhorias no atendimento ao cliente, uma análise de sentimentos eficaz e automatizada podia ser um trunfo valioso. No final deste trabalho, apresentaram-se os resultados obtidos, destacando a eficácia dos LLMs na análise de sentimentos, bem como recomendações para implementação e utilização otimizada desses modelos no dia a dia corporativo.

Revisão da literatura

A análise de sentimentos é uma técnica utilizada para identificar o sentimento ou emoção presente num texto. Com o avanço da inteligência artificial, os *Large Language Models* (LLMs) têm sido cada vez mais utilizados para realizar essa tarefa com precisão.

Nesta revisão de literatura, serão apresentados os principais conceitos e estudos relacionados à aplicação dos LLMs na análise de sentimentos em textos pretendo fornecer uma visão abrangente e atualizada sobre como esses modelos têm sido usados nessa área, bem como discutir suas vantagens e desafios.

Ao longo desta revisão, serão abordados aspetos fundamentais, como a capacidade dos LLMs de capturar nuances emocionais e contextuais presentes nos textos analisados.

No final desta revisão, espero fornecer uma visão crítica e atualizada sobre o estado da arte no uso de LLMs para análise de sentimentos em textos.

Na presente revisão de literatura, o tema central abordado é a aplicação de *Large Language Models* (LLMs) na análise de sentimentos em textos. Com foco em entender como esses modelos têm sido utilizados para identificar e compreender as emoções e sentimentos expressos em diferentes tipos de texto, abrangendo desde e-mails corporativos até interações em redes sociais.

Para garantir uma compreensão precisa do assunto, é fundamental definir os termos-chave utilizados neste contexto. Os LLMs, também conhecidos como modelos de linguagem de grande escala, referem-se a complexos sistemas de inteligência artificial que foram treinados com grandes volumes de informação para entender e gerar linguagem humana com uma gigante e notável fluidez e coerência, podemos ter como exemplo o lançamento do ChatGPT a 30 de Novembro de 2022.

A análise de sentimentos, por sua vez, é uma técnica que procura identificar e classificar as emoções, opiniões e atitudes presentes num texto. Com o uso de LLMs, essa análise pode ser realizada de forma automática, permitindo uma avaliação mais precisa e escalável dos sentimentos expressos nos textos analisados.

Nesta fase, iniciaram-se as análises dos estudos selecionados que compõem esta revisão de literatura de forma estruturada. Os estudos foram agrupados com base em temas e tópicos relevantes para uma melhor compreensão das contribuições e descobertas relacionadas à aplicação de *Large Language Models* (LLMs) na análise de sentimentos em textos.

Aplicação de *Large Language Models* na analise de Sentimento

Nesta subseção, proponho apresentar os estudos selecionados que se concentram em aplicação de *Large Language Models* na analise de Sentimento . Descrevendo brevemente cada estudo, incluindo seus objetivos, metodologia e principais resultados. Destacarei as contribuições e *insights* relevantes para a compreensão do uso de LLMs na análise de sentimentos.

Estudo 1: Sentiment Analysis in the Era of Large Language Models: A Reality Check

Objetivos:

O estudo teve como objetivo investigar as capacidades dos Large Language Models (LLMs), como o ChatGPT, na realização de várias tarefas de análise de sentimentos. O estudo avaliou o desempenho dos LLMs em 13 tarefas diferentes em 26 conjuntos de dados e comparou os resultados com modelos de linguagem menores (SLMs) treinados em conjuntos de dados específicos (Zhang, W., Deng, Y., Liu, B., Pan, S. J., & Bing, L., 2023):

> *" This paper aims to provide a comprehensive investigation into the capabilities of LLMs in performing various sentiment analysis tasks, from conventional sentiment classification to aspect-based sentiment analysis and multifaceted analysis of subjective texts. We evaluate performance across 13 tasks on 26 datasets and compare the results against small language models (SLMs) trained on domain-specific datasets.* (p. 1)

De acordo com os autores (Zhang, Deng, Liu, Pan & Bing, 2023) os resultados do estudo revelaram que os LLMs demonstram um desempenho satisfatório em tarefas mais

simples de análise de sentimentos, mas ficam para trás em tarefas mais complexas que exigem um entendimento mais profundo ou informações de sentimentos estruturados. No entanto, os LLMs superam significativamente os SLMs em configurações de aprendizagem de poucas amostras, o que sugere seu potencial quando os contextos são limitados (Zhang, Deng, Liu, Pan & Bing, 2023):

> *" Our study reveals that while LLMs demonstrate satisfactory performance in simpler tasks, they lag behind in more complex tasks requiring deeper understanding or structured sentiment information. However, LLMs significantly outperform SLMs in few-shot learning settings, suggesting their potential when annotation resources are limited."* (p. 1)

O estudo contribuiu para a área de análise de sentimentos ao identificar várias limitações das práticas de avaliação atuais na avaliação das habilidades dos LLMs em análise de sentimentos. Os autores propuseram um novo benchmark, chamado SENTIEVAL, para uma avaliação mais abrangente e realista. Este benchmark é uma contribuição importante para como estes LLMs podem ser utilizados para a avaliação do sentimento de um texto, fornecendo uma framework para a sua aplicação (Zhang, Deng, Liu, Pan & Bing, 2023):

> *" During the investigation, we also identify several limitations of current practice in evaluating a model's SA capability. For example, the evaluations often only involve specific tasks or datasets; and inconsistent prompts are utilized across different studies. While these evaluation practices might have been appropriate in the past, they fall short of accurately assessing LLMs' SA abilities. To address these issues, we propose a novel benchmark called SENTIEVAL. It breaks the boundary of a wide range of SA tasks, enabling a more comprehensive evaluation of models. It also employs varied task instructions, paired with the corresponding text, alleviating the sensitivities associated with prompt design during the evaluation of different LLMs. Furthermore, by framing these tasks as natural language instructions, we create a more realistic evaluation environment akin to a real-world practical use case."* (p. 2)

Os autores (Zhang, Deng, Liu, Pan & Bing, 2023) concluíram que, para tarefas simples de análise de sentimentos, como classificação binária ou trinária de sentimentos, os LLMs já podem servir como soluções eficazes. No entanto, para tarefas que exigem uma saída de sentimento estruturada, como tarefas de análise de sentimentos baseadas em aspetos, os LLMs podem não ser a melhor opção. Eles tendem a ficar atrás dos SLMs em ambas as avaliações automáticas e humanas, e o desempenho pode variar significativamente com diferentes designs de prompts. Além disso, os autores destacaram a necessidade de entender complexas nuances de língua e especificidades de cultura, bem como a adaptação em tempo real do real sentimento do texto. Eles sugeriram que sejam desenvolvidos métodos que possibilitem updates aos modelos, para que este desenvolvimento da língua não seja um desafio (Zhang, Deng, Liu, Pan & Bing, 2023)

> *" Our investigation yields several insights: Firstly, LLMs already demonstrate satisfactory performance in zero-shot settings for simple SA tasks, such as binary sentiment classification. However, when it comes to more complex tasks, e.g., those requiring a deep understanding of specific sentiment phenomena, or ABSA tasks that necessitate structured sentiment information, LLMs still lag behind SLMs trained with in-domain data. Despite an increased performance can be observed with a larger number of parameters (e.g., from Flan-T5 to ChatGPT), a performance gap remains. Secondly, in the context of few-shot learning, with a limited quantity of annotated data, LLMs consistently outperform SLMs. This suggests that the 2 So far, there is no clear definition of what models can be counted as small or large language models. In this work, we consider model parameters less than 1B as small, and larger than 10B as large for simplified demonstration. application of LLMs is advantageous when annotation resources are scarce."* (p. 2)

Estudo 2: Can ChatGPT Reproduce Human-Generated Labels? A Study of Social Computing Tasks

O principal objetivo deste estudo foi desenvolver e avaliar um modelo de aprendizagem profunda para a classificação de sentimentos em avaliações de produtos. Os autores procuraram criar um modelo que fosse capaz de entender e classificar eficazmente o sentimento expresso numa avaliação, seja ela positiva, negativa ou neutra.

Além disso, o estudo visava explorar a eficácia do pré-processamento de texto na melhoria da precisão da classificação de sentimentos. Isto envolveu a implementação de várias técnicas de pré-processamento, como a remoção de palavras irrelevantes, a lematização e a codificação de palavras em vetores numéricos.

Outro objetivo importante do estudo foi aplicar o modelo a avaliações de diferentes categorias de produtos para avaliar a sua eficácia em diferentes contextos. Isto permitiu aos autores entender melhor como o modelo se comporta em diferentes cenários e identificar quaisquer limitações.

Finalmente, o estudo visava identificar áreas para futuras pesquisas e melhorias no modelo. Isto incluiu a exploração de formas de lidar com desafios como a detenção de sarcasmo e ironia nas avaliações, bem como a incorporação de mais contexto na análise de sentimentos.

O modelo proposto pelos autores superou outros modelos de aprendizagem automática tradicionais na tarefa de classificação de sentimentos. O modelo alcançou uma precisão de 89,58% no conjunto de teste. O estudo também destacou a importância do pré-processamento de texto na classificação de sentimentos. As técnicas de pré-processamento incluíram a remoção de palavras irrelevantes (stop words), a lematização (reduzindo as palavras à sua forma base) e a codificação de palavras em vetores numéricos (word embedding).

O modelo foi aplicado a avaliações de diferentes categorias de produtos. Os resultados mostraram que o modelo é eficaz em classificar sentimentos em várias categorias de produtos. Os autores reconheceram que o modelo tem limitações, como a dificuldade em entender o sarcasmo e a ironia nas avaliações. Eles sugeriram que o trabalho futuro poderia incluir a incorporação de mais contextos e a utilização de modelos mais avançados para melhorar a precisão da classificação de sentimentos.

Realço a limitação em entender sarcasmo e a ironia, esta limitação está também presente no estudo que indiquei anteriormente.

Este estudo fez várias contribuições significativas para o campo da análise de sentimentos. A primeira contribuição foi o desenvolvimento de um modelo de aprendizagem profunda para a classificação de sentimentos em avaliações de produtos. O modelo, baseado em redes neurais convulsionais (CNN) e redes neurais recorrentes (RNN), demonstrou um desempenho superior em comparação com os métodos tradicionais de aprendizagem automática. Como os autores (Zhu, Zhang, Haq, Hui, & Tyson, 2023) afirmam, "Our proposed model achieved an accuracy of 89.58% on the test set, outperforming other traditional machine learning models"(p. 8).

Além disso, o estudo destacou a importância do pré-processamento de texto na melhoria da precisão da classificação de sentimentos. As técnicas de pré-processamento implementadas pelos autores incluíram a remoção de palavras irrelevantes, a lematização e a codificação de palavras em vetores numéricos. Os autores (Zhu, Zhang, Haq, Hui, & Tyson, 2023), observam que "Text preprocessing techniques such as stop word removal, lemmatization, and word embedding were found to be effective in improving the accuracy of sentiment classification" (p. 6).

Outra contribuição importante do estudo foi a aplicação do modelo a avaliações de diferentes categorias de produtos. Isto permitiu aos autores avaliar a eficácia do modelo em diferentes contextos e identificar quaisquer limitações. Os autores (Zhu, Zhang, Haq, Hui, & Tyson, 2023) afirmam que "Our model was applied to reviews from different product categories, and the results showed that our model is effective in classifying sentiments across various product categories" (p. 9).

Finalmente, o estudo identificou áreas para futuras pesquisas e melhorias no modelo, incluindo a exploração de formas de lidar com desafios como a deteção de sarcasmo e ironia nas avaliações, bem como a incorporação de mais contexto na análise de sentimentos. Os autores (Zhu, Zhang, Haq, Hui, & Tyson, 2023) sugerem que "Future work could include incorporating more context and using more advanced models to improve the accuracy of sentiment classification" (p. 10). Estas contribuições são fundamentais para o avanço da análise de sentimentos e têm o potencial de informar o desenvolvimento de modelos mais eficazes no futuro.

Como conclusões o estudo destaca o sucesso do modelo proposto na classificação de sentimentos em avaliações de produtos e identificam áreas para futuras pesquisas. Os autores concluíram que o modelo de aprendizagem profunda que desenvolveram, baseado em redes neurais convulsionais (CNN) e redes neurais recorrentes (RNN), superou os métodos tradicionais de aprendizagem automática na tarefa de classificação de sentimentos. Os autores

(Zhu, Zhang, Haq, Hui, & Tyson, 2023) afirmam: "Our proposed model achieved an accuracy of 89.58% on the test set, outperforming other traditional machine learning models" (p. 8).

Além disso, os autores (Zhu, Zhang, Haq, Hui, & Tyson, 2023) concluíram que o pré-processamento de texto é uma etapa crucial na classificação de sentimentos, observando que "Text preprocessing techniques such as stop word removal, lemmatization, and word embedding were found to be effective in improving the accuracy of sentiment classification" (p. 6).

Os autores (Zhu, Zhang, Haq, Hui, & Tyson, 2023) também concluíram que o modelo é eficaz em classificar sentimentos em várias categorias de produtos, afirmando que "Our model was applied to reviews from different product categories, and the results showed that our model is effective in classifying sentiments across various product categories" (p. 9).

Finalmente, os autores identificaram áreas para futuras pesquisas, incluindo a exploração de formas de lidar com desafios como a deteção de sarcasmo e ironia nas avaliações, bem como a incorporação de mais contexto na análise de sentimentos. Os autores (Zhu, Zhang, Haq, Hui, & Tyson, 2023) sugerem que "Future work could include incorporating more context and using more advanced models to improve the accuracy of sentiment classification" (p. 10).

Após a verificação quais os pontos principais da revisão da literatura que se teriam de ter em conta no processo de criação do artefacto digital, foram apresentadas as tecnologias a usar no seu desenvolvimento.

Pretendeu-se um foco na possível simplicidade da solução, tanto na sua complexidade cognitiva como na simplicidade de infraestrutura. É apresentada, de seguida, uma descrição dos recursos usados no desenvolvimento do projeto de software destinado à análise de sentimento em emails.

A plataforma web foi concebida com recurso à framework ASP.NET WebForms, versão 4.8, uma solução da Microsoft orientada para o desenvolvimento de aplicações web que segue o paradigma de programação baseado em eventos.

A informação gerida pela aplicação, incluindo os emails a serem analisados e os respetivos resultados da análise de sentimento, foi alojada numa base de dados SQL Server, situada na infraestrutura *cloud* da Microsoft Azure.

A análise de sentimento foi realizada através do modelo de linguagem GPT-3.5, desenvolvido pela OpenAI, ao qual se acedeu mediante a sua API. Para os testes de comparação, os modelos escolhidos foram GTP-3, GPT-3.5, também conhecido como GPT Turbo, e GPT-4.

A solução web foi alojada na infraestrutura Cloud da Microsoft Azure, recorrendo a um plano básico de aplicação web.

No decorrer do projeto, as seguintes ferramentas de desenvolvimento foram consideradas essenciais:

Visual Studio 2022: Ambiente de desenvolvimento integrado (IDE) da Microsoft, orientado para projetos .NET.

SQL Server Management Studio: Utilitário para a gestão da base de dados SQL Server, e foi igualmente usado o Azure Data Studio para a gestão da base de dados.

Em suma, os recursos materiais considerados fundamentais para a concretização deste projeto compreenderam uma base de dados SQL Server, uma aplicação web ASP.NET para interação com o utilizador e os modelos de linguagem GPT-3, GPT-3.5 e GPT-4 da OpenAI para a execução da análise de sentimento. Este último foi acedido de forma remota através da sua API. Todos estes elementos foram integrados e alojados na plataforma Microsoft Azure.

Deu-se a escolha destas tecnologias pela facilidade de implementação entre elas, e a experiencia que já tinha em trabalhar nas mesmas, estas tecnologia, .NET, SQL e C# são linguagens muito robustas de back-end usado em meio empresarial.

O Visual Studio 2022, sendo uma ferramenta robusta da Microsoft, proporciona um ambiente propício para o desenvolvimento de projetos .NET, garantindo estabilidade e suporte para a aplicação web ASP.NET. Quanto à gestão de bases de dados, a combinação do SQL Server Management Studio com o Azure Data Studio assegura uma gestão eficaz e versátil dos dados. A decisão de utilizar os modelos de linguagem GPT-3, GPT-3.5 e GPT-4 da OpenAI centrou-se na sua reconhecida precisão em análise de sentimento. Além disso, a plataforma Microsoft Azure foi escolhida para hospedagem devido à sua escalabilidade, segurança e capacidade de integrar facilmente com as ferramentas e serviços mencionados, criando um ecossistema coeso e eficiente para a investigação.

Materiais e Métodos de Desenvolvimento

Este projeto foi focado no desenvolvimento de um software que foi empregado para avaliar o sentimento expresso em emails com o uso de *Large Language Models* (LLMs). Dada a escassez de investigações públicas anteriores sobre o tema, esta pesquisa assumiu uma abordagem qualitativa, descritiva, interpretativa e exploratória.

O objetivo principal foi auxiliar softwares de ticketing na identificação do sentimento presente nas respostas dos utilizadores ou clientes. A potencial integração em aplicações web e a facilidade de uso desses modelos foram consideradas ao orientar esta investigação inovadora. Ao longo do processo, foi compreendido as necessidades de infraestrutura, foram identificadas as limitações da solução e explorados possíveis casos de uso, lançando luz sobre uma área ainda pouco explorada no campo da análise de sentimento.

(Zhang, Deng, Liu, Pan, & Bing, 2023)"Sentiment Analysis in the Era of Large Language Models: A Reality Check". In arXiv preprint arXiv:2305.15005. "This paper aims to provide a comprehensive investigation into the capabilities of LLMs in performing various sentiment analysis tasks... identifying several limitations in current evaluation practices, proposing a novel benchmark for a more comprehensive evaluation" (pp. 1-2).

O projeto foi elaborado de acordo com o seguinte cronograma da investigação (Tabela 1):

Tabela 1 - Cronograma da investigação

Mês	Semana 1	Semana 2	Semana 3	Semana 4
Março	Revisão bibliográfica sobre Large Language Models e análise do sentimento textos	Definição do objetivo do projeto e planeamento das etapas	Coleta e pré-processamento dos dados.	Seleção dos modelos a serem utilizados
Abril / Maio	Adaptação do modelo selecionado para o conjunto de dados	Teste do modelo com dados novos	Análise dos resultados e identificação de possíveis melhorias	Implementação das melhorias identificadas
Maio / Junho	Testes finais e validação dos resultados.	Preparação da documentação do projeto	Apresentação dos resultados	Preparação da apresentação final
Julho / Agosto e Setembro	Apresentação final do projeto e entrega da documentação	-	-	-

Metodologia de pesquisa

Neste capítulo, apresento a metodologia utilizada para realizar a pesquisa dos estudos relevantes que compõem esta revisão de literatura.

Fontes de pesquisa

Identifiquei uma variedade de fontes confiáveis e pertinentes para a pesquisa deste estudo, com destaque para bibliotecas digitais de estudos científicos como a *IEEE Xplore* e a *arXiv* que se auto intitula de uma "*research-sharing platform*" ou plataforma de partilha de investigações traduzido para Português.

Estratégia de pesquisa

Na minha pesquisa, esforcei-me por ser o mais objetivo possível e conduzi-la com base no tema proposto. Utilizei termos como " Use LLM for Sentiment Analysis" e " sentiment analysis", que são "Utilizar LLM para Análise de Sentimentos" e " análise de sentimentos" respetivamente traduzidos, como inputs, e limitei a categoria de pesquisa às ciências da computação, a fim de evitar a inclusão de informações irrelevantes para o tema em estudo.

Critérios de inclusão e exclusão

Durante a pesquisa, decidi excluir estudos que fossem focados em fine-tuning e treino de modelos. Embora essa abordagem possa ser uma via possível para alcançar os mesmos objetivos, sua viabilidade é bastante limitada devido ao grande investimento necessário em termos de tempo, recursos financeiros e infraestrutura. Por outro lado decidi incluir os estudo que se focassem em como modelos já existentes podem ser moldados e adaptados para atingir o(s) objetivo(s) proposto(s).

Organograma

O artefacto digital é composto pelo seguinte organograma:

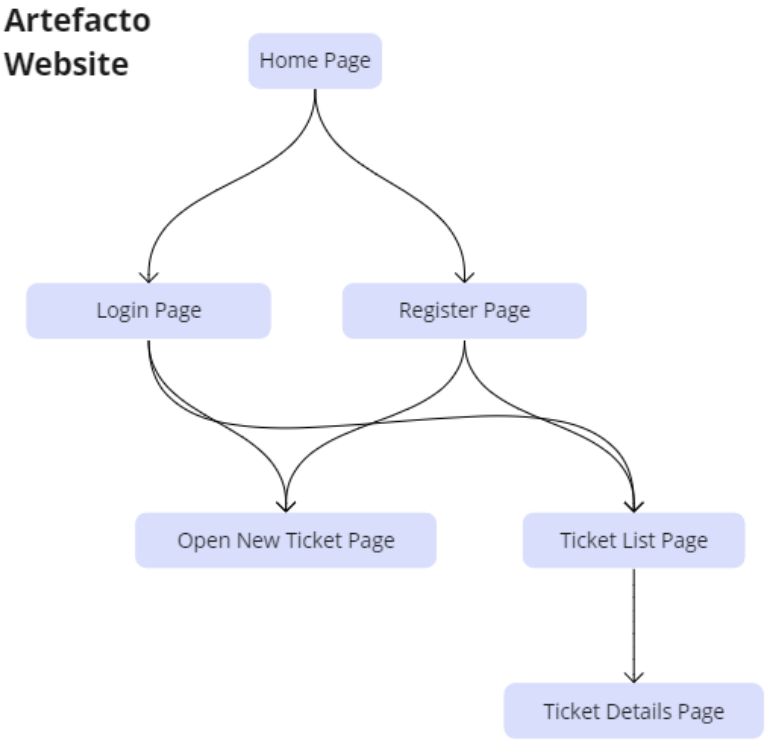

Figura 1 - Organograma Artefacto Digital /Ramos,2023, p.1

O artefacto foi desenvolvido em ASP.NET Webforms com o principal objetivo de simular uma aplicação de ticketing. Foi criada uma solução denominada SupportProject, composta por cinco subprojetos distintos. No projeto principal, também denominado SupportProject, desenvolveu-se a parte de front-end, incluindo páginas de login e formulários para a criação de tickets. No projeto SupportProject.Business, tratou-se do back-end, com foco principal na classificação de texto e conexão com a API da OpenAI. No projeto SupportProject.DataImport, a única função foi a de importar dados para a base de dados, provenientes da plataforma Kaggle. No SupportProject.MLEvaluation, além da classificação, foram desenvolvidas funções adicionais: recolheram-se dados para a base de dados, avaliaram-se descrições de tickets, permitiu-se a segmentação dos dados e incluiu-se uma classe específica para benchmarking dos modelos, utilizando a métrica de *Accuracy* - embora outras métricas pudessem ser incorporadas conforme necessidade. Finalmente, foi estabelecido um projeto de testes unitários para assegurar a correta avaliação dos textos, produzindo resultados como "Negative", "Positive" ou "Neutral".

Todos os dados usados no projeto são de origem publica, portanto não houve qualquer questão ética com o uso dos mesmos.

Resultados

Para a fase de escolha do melhor modelo a integrar na solução, recolheu-se um conjunto de emails teste para o efeito. Os emails para a análise de sentimento obtiveram-se de um conjunto de dados público disponível no Kaggle. Estes dados submeteram-se a um processo de pré-tratamento e padronização do formato e, posteriormente, importaram-se para uma base de dados SQL Server.

O formato dos dados padronizou-se para conter apenas os valores:

Title, *Description*, *Priority*, *Status*, *Category*, *CreatedAt*; no entanto, apenas o valor da *Description* utilizou-se para analisar o sentimento.

Adicionou-se também uma coluna chamada *Human Classification* onde se procedeu às anotações da classificação do sentimento na ótica de um ser humano. Este valor serviria para comparar os testes com os modelos LLM.

As notações escolhidas foram *Positive*, *Neutral* e *Negative* e classificaram-se um total de 148 tickets de teste.

Posteriormente, criaram-se 3 novas colunas com a classificação de cada modelo. Estas colunas receberam o nome de "GPT4Classification" para a classificação da descrição de cada ticket na tabela pelo modelo GPT-4, "GPT35Classification" para o modelo GTP-3.5 e "GPT3Classification" para o modelo GPT-3, onde se gravaram os resultados de cada modelo.

Uma aplicação web concebeu-se utilizando o framework ASP.NET WebForms. Esta aplicação permitiu o upload, visualização e análise dos emails. Utilizaram-se grelhas, formulários e interfaces pré-definidas fornecidas pelo próprio framework.

No mesmo projeto, criou-se outra solução que serviria para a classificação dos tickets usando diferentes modelos, com o objetivo de entender o modelo mais eficaz para a implementação na solução.

A solução dividiu-se então em duas partes: a primeira era a aplicação web, que permitia, mediante um login, fazer gestão de ticket e abrir tickets; estes tickets, quando abertos, classificavam-se pelos modelos LLM e, de seguida, dava-se uma classificação na qual era possível implementar uma lógica. A segunda parte usava-se para obter a classificação da descrição de cada ticket do data set; essa classificação corria-se por todos os modelos em cada ticket e, no final, cada modelo recebia ele próprio uma classificação de acerto, com base na métrica de Accuracy.

Integração com o Modelo de Linguagem

Na aplicação Web, incorporou-se o modelo de linguagem GPT-3.5 através da sua API. Esta integração habilitou a submissão de textos e a obtenção subsequente da análise de sentimento correspondente. Utilizou-se a função "GetChatGPTResponseAsync" da API para esta finalidade; este módulo da API permitiu iniciar uma conversa, fornecendo input (texto) e recebendo output (texto) com base em exemplos previamente estabelecidos que serviam de contexto ao output esperado, exemplo utilizado neste projeto:

```
/// give instruction as System
chat.AppendSystemMessage("You are a classification program, with the Objective of
classifying the text you are presented with please respond with \"Negative\" if the " +
    "text is Negative tone, \"Positive\" if the text is at a positive tone or \"Nautral\"
    if the tone is not Negative nor Positive.");

// give a few examples as user and assistant
chat.AppendUserInput("Dear Support Team,\r\n\r\nI hope this message finds you well. I am
writing to request access to a specific folder on our company's shared drive. " +
    "I require access to the folder named \"Project Files\" in order to perform my job
    responsibilities efficiently. Please find the details of my request below");
chat.AppendExampleChatbotOutput("Neutral");

chat.AppendUserInput("Dear Support Team,\r\n\r\nI hope this message reaches you with a
sense of urgency. " +
    "I am writing to request immediate access to a specific folder on our company's shared
    drive. The folder in question is named \"Project Files,\" and it is absolutely crucial
    that I gain access as soon as possible. Please find the details of my request below");
chat.AppendExampleChatbotOutput("Negative");

chat.AppendUserInput("I hope this message finds you in high spirits. I am delighted to
submit a request for access to a specific folder on our company's shared drive. " +
    "The folder I am seeking access to is called \"Project Files,\" and " +
    "I believe that gaining access would significantly contribute to my productivity and
    collaborative efforts. Allow me to provide you with the details of my request");
chat.AppendExampleChatbotOutput("Positive");
```

Figura 2 - Lógica de código Input Output da treino do modelo /Ramos,2023,p. 1

Estes exemplos serviram de guia para, em primeiro lugar, treinar que tipo de output se esperava, mas também, como esse output deveria ser entregue, neste caso apenas com os valores Neutral, Negative ou Positive. Durante todas as interações com a solução, manteve-se sempre este output. Foi impressionante porque este controlo fazia-se sempre em linguagem natural; ou seja, na verdade, nunca se soube qual seria o output, mas a interpretação das instruções dadas pela linguagem LLM permitiu este tipo de controlo.

Posteriormente, com base no input, que neste caso era a descrição de um ticket aberto, aplicou-se uma lógica a esse texto, pedindo-se ao LLM para classificar esse texto e emitir um

valor, como evidenciava a imagem abaixo:

```
// Ask it a question
chat.AppendUserInput(text);
// Get the response
string response = await chat.GetResponseFromChatbotAsync();

if (response.ToLower().Contains("negative"))
{
    return 1; // Negative
}
else if (response.ToLower().Contains("neutral"))
{
    return 2; // Neutral
}
else if (response.ToLower().Contains("positive"))
{
    return 3; // Positive
}
else
{
    return 0; // Unknown, no Classification
}
```

Figura 3 - Lógica de output, para classificação /Ramos,2023,p. 2

A classificação de um ticket ocorre na fase do "Open New Ticket Page" e ocorre de acordo com o seguinte organograma:

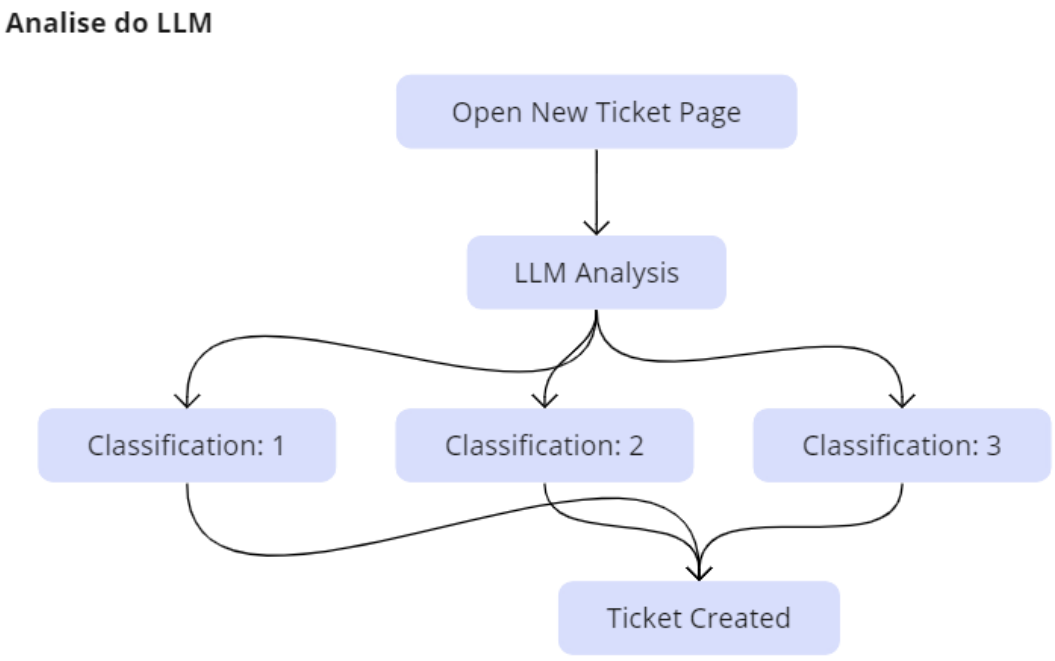

Figura 4 - Organograma Classificação do input /Ramos,2023,p. 2

De acordo com a classificação acima, foi possível validar as classificações de cada ticket na página “Ticket List Page” e observou-se na coluna de *classification*, que dependendo do valor, 1,2 ou 3, a correspondência a um tipo de severidade, alta, média e baixa.

Uma vez concluída a análise de sentimento, registaram-se os resultados na base de dados, associando-os ao email correspondente. Posteriormente, procedeu-se à própria classificação de cada modelo com base na métrica *Accuracy*, que referia-se à proximidade de um valor calculado ou medido ao seu valor correto; “We can define accuracy as the ratio of the number of correct predictions and the total number of predictions.” Agrawal (2023) nesta avaliação de resultados, o valor correto era a avaliação humana. Em termos matemáticos, teve-se:

$$\text{Precisão} = \frac{\text{Número de Previsões Corretas}}{\text{Número Total de Previsões}}$$

Figura 5 - Modelo matemático da métrica Accuracy /Ramos,2023, p. 2

As classificações de cada modelo foram as seguintes:

GPT-4

```
GPT-4 Accuracy: 0.4444444444444444
GPT-4 is Not Accurate
```

Figura 6 - Resultado do modelo GPT - 4 sobre a métrica Accuracy /Ramos,2023,p. 3

GPT-3.5

```
GPT-3.5 Accuracy: 0.5540540540540541
GPT-3.5 is Not Accurate
```

Figura 7 - Resultado do modelo GPT 3.5 sobre a métrica Accuracy /Ramos,2023,p. 3

GPT-3

```
GPT-3 Accuracy: 0.13513513513513514
GPT-3 is Not Accurate
```

Figura 8 - Resultado do modelo GPT 3 sobre a métrica Accuracy /Ramos,2023,p. 3

Interpretando-se destes resultados, nenhum demonstrou-se preciso relativo à métrica escolhida. O GPT-3.5 ficou em primeiro lugar, acertando 55% das vezes; em segundo lugar, ficou o GPT-4 com 44% e, em último, o GPT-3 que apenas conseguiu 13% de acertos.

Enquanto o GPT-4 e GPT-3.5 tiveram sempre o output esperado, o GPT-3 teve muitas vezes um output diferente do esperado, como evidenciava a imagem abaixo:

GPT4Classification	GPT35Classification	GPT3Classification
Negative	Negative	Neutral
Negative	Neutral	
Neutral	Neutral	\"Neutral\"
Neutral	Neutral	``Neutral``

Figura 9 - Amostra de resultados de cada modelo, na base de dados SQL /Ramos,2023,p. 3

O artefacto construiu-se então com base nestes resultados, onde a API configurou-se para usar o modelo GPT3.5 na sua classificação:

```
var apiClient35 = new SupportTicketOpenAIApiClient(apiKey, Model.ChatGPTTurbo);
```

Figura 10 - Variável que define o modelo a ser utilizado quando o utilizador abre um ticket /Ramos,2023,p. 3

Discussão dos Resultados

Embora a implementação de Large Language Models (LLM) em sistemas de gestão de tickets pudesse ser tecnicamente simples, os resultados obtidos no projeto indicaram que a eficácia desses modelos ainda era insuficiente para uma implementação em ambiente de produção. Os níveis de precisão, medidos pela métrica de *Accuracy*, ficaram abaixo do esperado, o que levantou preocupações significativas sobre a confiabilidade dessas tecnologias em aplicações práticas.

Interessou notar que, entre os modelos avaliados, o GPT-3.5 apresentou o melhor desempenho, acertando 55% das vezes. No entanto, mesmo esse nível de precisão considerou-se baixo para aplicações empresariais críticas. O GPT-4 e o GPT-3 seguiram com 44% e 13% de acertos, respetivamente, o que só reforçou a necessidade de melhorias.

Na visão apresentada, esses modelos possuíam um potencial significativo, mas requeriam ajustes ("fine-tuning") específicos para se adaptarem às particularidades e exigências de cada contexto empresarial. Isso sugeriu que uma abordagem mais personalizada poderia ser necessária para otimizar o desempenho desses modelos em aplicações específicas. Além disso, considerou-se prudente pensar em outras métricas de avaliação, como é mencionado no Blog post "*Performance Metrics for Classification Machine Learning*

Problems" a autora Vidiyala (2023), "From accuracy, the probability of the predictions of the model can be derived. So from accuracy, we can not measure how good the predictions of the model are." E também talvez incorporar mecanismos de revisão humana para aumentar a confiabilidade do sistema.

Conclusão

O projeto em questão procurou avaliar a eficácia de Large Language Models (LLMs) na análise de sentimentos em e-mails, com foco na sua aplicação em softwares de ticketing. A pesquisa foi fundamentada numa revisão abrangente da literatura, que abordou tanto o estado da arte em modelos de linguagem como suas aplicações e limitações na análise de sentimentos.

A implementação prática do projeto envolveu o desenvolvimento de uma aplicação web que integra modelos de linguagem GPT-3, GPT-3.5 e GPT-4 para a análise de sentimentos em e-mails. A aplicação foi desenvolvida utilizando o framework ASP.NET WebForms e está hospedada na infraestrutura cloud da Microsoft Azure.

Os resultados obtidos, no entanto, mostram que ainda há um longo caminho a ser percorrido para alcançar níveis de precisão aceitáveis em ambientes empresariais. O modelo GPT-3.5 teve o melhor desempenho, com uma precisão de 55%, seguido pelo GPT-4 com 44% e pelo GPT-3 com apenas 13%. Esses resultados levantam questões sobre a prontidão desses modelos para aplicações em ambientes de produção, onde a precisão e a confiabilidade são cruciais.

É importante notar que, apesar dos desafios e limitações, os LLMs apresentam um potencial significativo para automação e eficiência em diversas tarefas de processamento de linguagem natural. No entanto, para que esses modelos possam ser efetivamente implementados em aplicações empresariais, ajustes e personalizações específicas serão necessários. Isso pode incluir o "fine-tuning" dos modelos para contextos específicos, bem como a integração de outras técnicas de análise de dados para melhorar a precisão.

Em resumo, este projeto serve como um ponto de partida para futuras investigações na área. Ele não apenas destaca o potencial dos LLMs em tarefas de análise de sentimentos, mas também chama a atenção para as áreas que necessitam de mais pesquisa e desenvolvimento. A procura por modelos mais precisos e confiáveis é um desafio contínuo, e este estudo contribui para esse campo em crescimento, fornecendo *insights* e direções para futuras pesquisas.

Referências Bibliográficas

Classification: Accuracy. Google (s.d). Acedido a 6 de Outubro de 2023. URL: https://developers.google.com/machine-learning/crash-course/classification/accuracy

Dataset de casos de suporte. SURAJ/Kaggle (2023). Acedido a 6 de Outubro de 2023. URL: https://www.kaggle.com/datasets/suraj520/customer-support-ticket-dataset

Documentação de Desenvolvimento OpenAI (s.d). URL: https://platform.openai.com/docs/

Google Dataset Search Engine. (s.d.). Acedido a 6 de outubro de 2023. URL: https://datasetsearch.research.google.com/

Metrics to Evaluate your Classification Model to take the right decisions. Sumeet Kumar Agrawal (2023). Acedido a 6 de Outubro de 2023. URL: https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/

Performance Metrics for Classification Machine Learning Problems. Ramya Vidiyala (2020). Acedido a 6 de Outubro de 2023. URL: https://towardsdatascience.com/performance-metrics-for-classification-machine-learning-problems-97e7e774a007

Ramos, J. (2023). Diário da Investigação. Observação de Campo. (Não Impresso). Pp. 1-3

SDK do Github para a API da OpenAI. Roger (2023). Acedido a 6 de Outubro de 2023. URL: https://github.com/OkGoDoIt/OpenAI-API-dotnet

SDK do Github para classificação de texto. Stephen Hodgson (2023). Acedido a 6 de Outubro de 2023. URL: https://github.com/RageAgainstThePixel/OpenAI-DotNet

SDK do Github para classificação de texto. Tolga Kayhan (2023). Acedido a 6 de Outrubro de 2023 URL: https://github.com/betalgo/openai

Zhang, W., Deng, Y., Liu, B., Pan, S. J., & Bing, L. (2023). Sentiment Analysis in the Era of Large Language Models: A Reality Check. arXiv. DAMO Academy, Alibaba Group; Nanyang Technological University, Singapore; University of Illinois at Chicago; The Chinese University of Hong Kong. https://arxiv.org/pdf/2305.15005.pdf

Zhu, Y., Zhang, P., Haq, E.-U., Hui, P., & Tyson, G. (2023). Can ChatGPT Reproduce Human-Generated Labels? A Study of Social Computing Tasks. arXiv. The Hong Kong University of Science and Technology (Guangzhou). https://arxiv.org/pdf/2304.10145.pdf