Natural Language Processing and its applications in e-business

Gabriela Inês Carneiro de Sousa

ISCAP, P.PORTO

Abstract: The advent of big data and the ability to extract insights from unstructured data has opened new avenues for companies. In this scientific article we have gathered different examples of application of Natural Language Processing (NLP) specifically in the electronic business environment. The main objective was to investigate what kind of analyzes and techniques are being used to extract knowledge from Big data, more concretly unstructured data, and within this type, textual data and what concrete applications derive from NLP in the context of e-business. Being NLP a type of Artificial Intelligence (AI) that uses Machine Learning (ML) and Deep Learning (DL) with the objective of developing a technology that learns, and takes decisions based on what it learned, it was important to understand what a ML workflow is like and therefore, the literatuere revision is very focused on this point. It was concluded that NLP is a complex process whose applications are diverse and relevant in the context of Electronic Business. Today, NLP is used to complete tasks such as text classification, content filtering, sentiment analysis, language modeling, translation, and summarization and applications such as chatbots, voice assistants and recommendation systems. In the future, it is expected that the NLP will have greater influence in different business areas such as the management of new products and recommendation systems; segmentation, segmentation and analysis of customers and users; brand positioning, communication and marketing; competitor analysis; and risk management, sustainability and social responsibility. Specialists in the area also believe that there will possibly be a paradigm shift in AI that will use Reinforcement Learning techniques, which will allow the development of more advanced, adaptive and multipurpose AI agents. It is also to be expected that humans and machines cohabit in a more collaborative way.

Key words: E-business, Natural Language Processing (NLP), Unstructured Data, Big data analytics, Machine Learning (ML)

Resumo: O advento da big data e a capacidade de extrair *insights* de dados não estruturados abriram novos caminhos para as empresas. Neste artigo científico reunimos diferentes exemplos de aplicação de Processamento de Linguagem Natural (PLN) em específico no ambiente de negócios eletrónicos. O objetivo principal foi investigar que tipo de análises e técnicas estão a ser utilizadas para extrair conhecimento de Big Data, mais concretamente de dados não estruturados, e dentro deste tipo, dados textuais e quais aplicações concretas derivam da PNL no contexto de e-business. Sendo a PLN um tipo de Inteligência Artificial (IA) que utiliza Machine Learning (ML) e Deep Learning (DL)

com o objetivo de desenvolver uma tecnologia que aprende, e toma decisões com base no que aprendeu, foi importante entender o que é o fluxo de trabalho de um projeto de ML e, portanto, a revisão da literatura está muito focada nesse ponto. Concluiu-se que o PLN é um processo complexo cujas aplicações são diversas e relevantes no contexto dos Negócios Eletrónicos. Hoje em dia, o PLN é utilizado para completar tarefas como classificação de texto, filtração de conteúdo, análise se sentimento, modelação de língua, tradução, e sumarização e aplicações como chatbots, assistentes de voz e sistemas de recomendação. No futuro, é expectável que o PLN tenha maior influência em distintas áreas de negócio como a gestão de novos produtos e sistemas de recomendação; segmentação, segmentação e análise de clientes e usuários; posicionamento, comunicação e marketing da marca; análise de concorrentes; e gestão de riscos, sustentabilidade e responsabilidade social. Especialistas da área acreditam também que haverá possivelmente uma mudança de paradigma no AI que passará a utilizar técnicas de Reinforcement Learning o que permitirá o desenvolvimento de agentes de AI mais avançados, adaptativos e polivalentes. É de esperar também que humanos e máquinas coabitem de forma mais colaborativa.

Introduction

Until recently, most data collected by organizations consisted of structured transaction data that could easily fit into rows and columns of relational database management systems. Since then, there has been an explosion of data from web traffic, e-mail messages, social media content (tweets, status messages), even music playlists, as well as machine-generated data from sensors that, due to the plummeting costs of data storage and powerful new processing capabilities, can now be stored and analyzed to draw connections and make inferences and predictions. This data may be unstructured or semi-structured and thus not suitable for relational databases structures. The term big data refers to this avalanche of digital data that creates huge data sets, often from different sources, in the petabyte and exabyte range. (Laudon & Traver, 2021).

Some examples of big data are analyzing 8 terabytes of tweets generated by Twitter each day to improve your understanding of consumer sentiment toward a product; 100 million e-mails in order to place appropriate ads alongside the e-mail messages; 500 million call detail records to find patterns of fraud and churn (Laudon & Traver, 2021).

NLP falls, therefore, within the umbrella of Big Data. NLP involves a range of computational techniques useful for analyzing human communication in different languages. These methods have gained prominence with the spurt in computational power, analytical abilities, and processing speed. Thanks to these developments and advances in its parent field of machine learning NLP has achieved several milestones (Shankar & Parsana, 2022), such as powerful chatbots and voice assistants.

NLP is about teaching machines to replicate the human way of communicating, which covers both written and spoken language. NLP is a critical aspect in the development of Artificial Intelligence (AI), especially for one requiring language inputs, such as in voice assistants (Kotler et al, 2021). However, deeper concepts guide AI development today than merely parroting of human language via computers. AI research looks to develop technology that takes action based on learned patterns (iPullRank, 2017).

It is also a challenging feat since the human language, in its natural form, is often fuzzy, intricate, and layered. A large volume of real conversation transcripts and video recording is required to teach machines the nuances of language (Kotler et al, 2021).

The objectives of this work are to identify the current and potential future usages of NLP in the context of e-business. For the purpose, the article's structure starts with a literature review, followed by the methodology, the results, a discussion section and the conclusion.

Literature Review

In computer science, one refers to human languages, like English, Portuguese or Mandarin, as "natural" languages to distinguish them from languages that were designed for machines, where rules came first, and people only started using the language once the rule set was complete. With human language, it's the reverse. Natural language was shaped by an evolution process. Its "rules", like the grammar of English, were formulized after the fact and are often ignored or broken by its users as a result, while machine-readable language is highly structured and rigorous, using precise syntactic rules to weave together exactly defined concepts from a fixed vocabulary. NLP is about using machine learning and large datasets to give computers the ability to ingest a piece of language as input and return something useful (Chollet, 2021).

The overall machine learning workflow include the following steps:

1) Gathering and collecting the relevant data for your task: Researchers in business have used text data from sources as online browsing, conversations, firm communications, video and audio files, and social media chatter to perform NLP tasks (Shankar & Parsana, 2022).

2) Cleaning and inspecting the data to better understand it: Several pre-processing methods can be used, but some of the most common involve pre-processing to remove unwanted words, numbers, punctuation and/or symbols (Shankar and Parsana, 2022)..

3) Performing feature engineering to allow the algorithm to leverage the data in a suitable form (e.g. converting the data to numerical vectors). Before being able to run an NLP model, the data, which are usually in the form of human readable texts, should be converted into a well-ordered input suitable for analysis by algorithms (Chambers and Zaharia, 2018).

4) Split the data into train and test and try different models: Statistical NLP models are typically text classification models based on statistical distributions of words. A topic model is a commonly used model that extracts keywords grouped together with an underlying thematic similarity or topic. Neural NLP models use deep learning. Deep learning uses the human brain as its design model. It layers brain-like neurons created from levels of machine learning. Each level does its learning and produces results that get passed onto the next network that takes on another task with that data. (iPullRank, 2017).

5) Evaluating the model: Different NLP models use different methods for classification and prediction. We can compare and evaluate NLP models on different metrics. Because most models rely on classification tasks, metrics like accuracy, precision, recall, error rate are appropriate metrics (Shankar and Parsana, 2022).

Although algorithms are usually automated once in operation, their development, installation, and training remain highly technical, research-intensive, and human-centric activity. Humans play a strategic role in the ongoing fine-tuning of AI systems that lead to optimized processes. In fact, AI is not a “set-and-forget” technology as models are continuously tuned manually, especially when natural language processing is involved. (Mari, 2019).

Ivo Nogueira and Isabel Chaves, both experienced Data Scientists, have shared the same conviction of non-linerarity, non-fully-automatiton and human-centric activity

when asked if they agreed with the Machile Learning workflow abeforemencioned. Ivo responded: “I agree with the steps listed, although it should be noted that they are often done out of order and even multiple times for the same project. Lastly, every project should prepare for the point in time where drift is going to make its results obsolete.”. Isabel, in its turn, stated: “Although I do agree with those steps, in most projects they are not in a linear order. Most of the times we need to go back and forth in order to obtain real progress.”

Methodology

The research method used to complete this research was to interview five professionals and researchers in the are of NLP, with the objective of cross-checking and validating the theoretical findings. I have put together a total of six questions as the interview guide (Apendix I) and received four replies out of the five invitations sent.

Results

NLP has profound use cases in business, in particular, eletronic and digital business. It is useful to business professionals for making strategic decisions and for improving existing products and services. The rise of mobile devices has enabled users to instantly post brief texts to express thoughts and emotions. Tapping this information is vital for enhancing customer experience, forecasting sales and formulating strategies. Other applications of NLP in e-business include customer credit assessment, social media monitoring, chatbots, and personal assistants (Shankar and Parsana, 2022).

Amongst the possible usages of NLP the we have Text Classification which aims to classify the topic of a text; Content Filtering that can for example determine if a text contains an abusive message; Sentiment analysis which pertains to judge the polarity of a text, if it conveys a positive ou negative message. Sentiment analysis or opinion mining is the process of identifying the feeling or emotion expressed in the text or document (Yarkareddy et al., 2022); Language modeling aims at predicting the next word in an incomplete sentence; Translation which ams to translate a tex from one language to another; and Summarization that has the objective of reducing the main ideas of a text or document to a version with less words (Chollet, 2021).

The most widespread application of NLP is for chatbots. Essentially, an NLP refers to components that allow chatbots to carry out a conversation that’s nearly identical to a

person-to-person human conversation. NLPs use deep learning to not only study human input but also generate a human response. According to Chatbots Magazine, newer, smarter chatbots using NLP should be able to carry out the following functions: Summarizing large quantities of text into concise explanations; Responding to open or closed questions ("Is Washington, D.C., the capital of the United States" compared to "Why is Washington, D.C., the capital of the United States?") , Inferring conference clues (that is, relating an object with words in a sentence), Ambiguity (sorting out the context and meanings of words in sentences), Morphology (the ability to separate individual words into morphemes, the smallest unit of meaning for words), Semantics (the meaning of sentences or words in natural human language), Text structure (the usage of punctuation, spacing, and text), Sentiment (the emotional "feeling" conveyed by the speech patterning, for example, "happy" or "sad") (iPullRank, 2017).

Chatbots will reduce the need for higher-cost channels such as inbound call centers and outbound marketing, especially when it comes to serving lower-tier customers.. The popularity of online messaging platforms, such as WhatsApp, Facebook Messenger, and WeChat are the key contributor to the rise of chatbots (Kotler et al, 2021). With 100,000 chatbots running on Facebook Messenger, companies are increasingly interacting with digital consumers to obtain leads, within the scope of E-Commerce and customer support (Cortez, 2018). For this reason, people expect to communicate with chatbots in the same way they would chat with other people (Kotler et al, 2021).

With voice tech, machines have also gotten much better at responding to verbal commands. There are many available voice assistants: Amazon Alexa, Apple Siri, Google Assistants, and Microsoft Cortana. These applications are already very capable of answering simple queries and executing commands in multiple languages (Shankar and Parsana, 2022). The latest AI-fueled platforms (assistants) can extract relevant pieces of information from both verbal and textual conversations in real-time to swiftly capture popular issues, suggest next best action to agents, and predict the likelihood of a customer to churn (Alex, 2019).

Salesforce's Einstein leverages rules-based and predictive models to provide agents with contextual recommendations and offers for customers. These "next best actions" suggested to employees, such as "give free shipping" or "offer zero percent financing" lead to higher customer loyalty and upselling opportunities (Alex, 2019).

A different approach to an application of NLP was investigated by Zoghbi et al. (2016) in their cross-modal search solution of fashion items. Given a textual query composed of visual attributes of dresses, the system retrieves relevant images of dresses, and given a picture of a dress as query, the system describes the attributes of the dress in natural language terms.

Shankar and Parsana (2022) also identify the potential future research directions for NLP models in e-business dividing them by different business objectives and tasks such as new product management, segmentation and targeting, brand position, crisis management, competitor analysis, Management of sustainability and corporate social responsibility, amongst others.

When asked what were the main applications of NLP in e-business today and what new applications we might expect to see in the future, Data Scientist Isabel Chaves replied respectively “Chatbots, but I think that we need to surpass the fear that talking to a machine might cause.” and “To be truly honest, I don't know. I hope that in the future we are able to introduce Machine Learning to the health systems. Not in a way to replace doctors but to help them (collaborative working), but there is still a lot of work to do since there is a lack of trust on what a machine can do.”

Data Scientist Ivo Nogueira has given “Recommender systems, machine translation and chat bots” as “probably the best/most profitable examples of the industry” when referring to the mains applications on NLP nowadays. For the future, he shared the opinion that Reinforcement Learning – a machine learning method based on rewarding desired behaviors and/or punishing undesired ones (Carew, n.d.) - is the way forward: “In the future I expect applications to use multiple fields of data science at the same time. These, along with tasks where the outputs change frequently, are the tasks where humans still significantly outperform machines. Currently the best approach for adaptable models seems to be reinforcement learning approaches, which is where most of the innovation is going to come from in the next decade.”

Data Scientist João Sá, in its turn has shared cost concerns in training the models and indicated that a way forward might be to make these models and techniques more accessible: “The training of these types of models are quite expensive, where we need to somehow index all text available into a Machine Learning model. So in the future I expect

that training these models will be easier in a way that we could evangelize these types of techniques to the whole world." Once this is achieved, he sees that a relevant new feature to develop, especially for chatbots and content creators, would be personalization: "Also most of these models don't consider personalization, I expect that chat bots in the future will be able to create text on demand and personalize to any user profile. Also, this type of text generation could help content creators by providing already recommendations of product descriptions, news, etc."

Discussion

Natural Language Processing (NLP) is the type of Artificial Intelligence which uses Machine Learning and Deep Learning techniques in order to have machines making sense and extracting something useful from textual data.

The current applications of NLP are quite clear and most of us already use them frequently on our daily lives: chatbots, voice assistants, recommendations systems, translating applications, etc. It is harder to understand the way in which it will develop in the future. It is expected that technology will evolve and all big data being collected by our interactions with digital environments will be used to feed systems and develop multipurpose adaptable agents that will have an impact in different domains of the electronic business. The ways the technology will develop is unclear though, and even experts are not quite certain of what kind of techniques may be used and which kind of human-machine relationship will emerge.

In my opinion, the confusion comes from the fact that concepts like Big Data, Artificial Intelligence, Machine Learning and Natural Language Processing, amongst others, are still quite ambiguous for the general public and even within the scientific community. It is important to continue having open discussions about these topics in order to better understand the direction they are taking us to and their implications in our present and future lives.

Conclusion

Natural Language Processing (NLP) is a form of Artificial Intelligence that gives machines the ability to read and interpret human language. With NLP, machines can make sense of written or spoken text.

To make this happen, NLP goes through a Machine Learning workflow that is divided in five phases: firstly text data is collected from online sources, videos or audio files; secondly data should inspected and cleaned recurring to text-specific techniques; in a third phase feauture engineer should be performed in order to convert the data into a suitable form that machines can read; the forth step is to use split the data into train and test data and run one or several different models; and the fifth and final step is to evaluate the models basing the evaluation on different metrics such as accuracy, precision, recall, etc, and choose the best amongst the candidates. These five steps may be done out of this order and even have to be performed multiple times for the same project. Human supervision is therefore fulcral in the ongoing fine-tunning of AI, and therefore NLP.

NLP is constantly evolving, but existing NLP-based solutions include text classification, content filtering, sentiment analysis, language modeling, translation, summatization, chatbots, voice assistants and recommendations systems.

There is reason to believe that its usage will further develop in the future and allow companies to perform analyzes based on their user generated text data such as comments on social media and reviews that will influence and help better understand and predict elements related to several business domains such as segmentation and targeting, brand position, crisis management, competitor analysis, and canagement of sustainability and corporate social responsibility. Experts in the area believe that the advancements will occur in the direction of reinforcement learning algorithms in which agents learn and improve through a reward and/or penalty based system, and of human–technology collaboration, in which, machines and programs do not replace humans in their activities but rather support them bringing more effenciency or enhancing these tasks. Additionally, it is given an emphasis to the feature of personalisation.

In terms of limitations, I wished I have had more time to dedicate to the reading of more scientific articles and be able to elaborate a thorough comparison of different authors and to explore the more technical and practical side of the subject by collecting data and developing a model myself. Nevertheless, I believe that the main objectives of the research were met: finding the current techniques to perform an NLP task and their current and potencial future applications.

References

Carew, J. (n.d). Definition: Reinforcement Learning. https://www.techtarget.com/searchenterpriseai/definition/reinforcement-learning

Chambers, B., Zahari, M. (2018). Spark The Definite Guide: Big Data Processing Made Simple, O'reilly

Chollet, F., (2021). Deep Learning with Python, Manning

Cortez, D. (2018). O uso de Chatbots em experiências de Mobile Commerce em Portugal, Dissertation ISCAP

iPullRank (2017). Machine Learning for Marketeers: A Comprehensive Guide to Machine Learning

Kotler, P., Kartajaya, H., Setiawan, I. (2021). Marketing 5.0: Technology for Humanity, Willey

Laudon, K. C., Traver, C. G. (2022.) E-commerce 2021-2022: Business. Technology. Society, Pearson

Mari A. (2019). The Rise of Machine Learning in Marketing: Goal, process, and benefit of AI-Driven Marketing, Swiss Cognitive

Shankar, V., Parsana, S. (2022). An overview and empirical comparison of natural language processing (NLP) models and an introduction to and empirical application of autoencoder models in marketing, Journal of the Academy of Marketing Science

Yarkareddy, S., Sasikala, T., Santhanalakshmi, S. (2022). Sentiment Analysis of Amazon Fine Food Reviews, Department of Computer Science and Engineering, Amrita School of Engineering

Zoghbi, S., Heyman, G., Gomez, J. G., Moens, M. (2016). Fashion Meets Computer Vision and NLP at e-Commerce Search, International Journal of Computer and Electrical Engineering

Appendices

Appendix I – Interview Questions

1) Could you please write your name, role and company you're working on currently?

2) Do you have working or academic experience with Natural Language Processing? If yes, could you develop on the project/tasks you were involved with?

3) In our research, we have identified the following steps in the modelling of an NLP problem: a. Gathering and collecting the relevant data for your task; b. Cleaning and inspecting the data to better understand it; c. Performing feature engineering to allow the algorithm to leverage the data is a suitable form (e.g. converting the data to numerical vectors); d. Using a portion of this data as a training set to train or or more algorithms to generate some candidate models. Do you agree?

4) Could you express your opinion together with a short description of a specific example where you also mention the tools/programs/software's used to execute the tasks?

5) What are in your opinion the main applications of NLP in e-business today?

6) What new applications do you expect to see in the future? In what do you base these expectations?