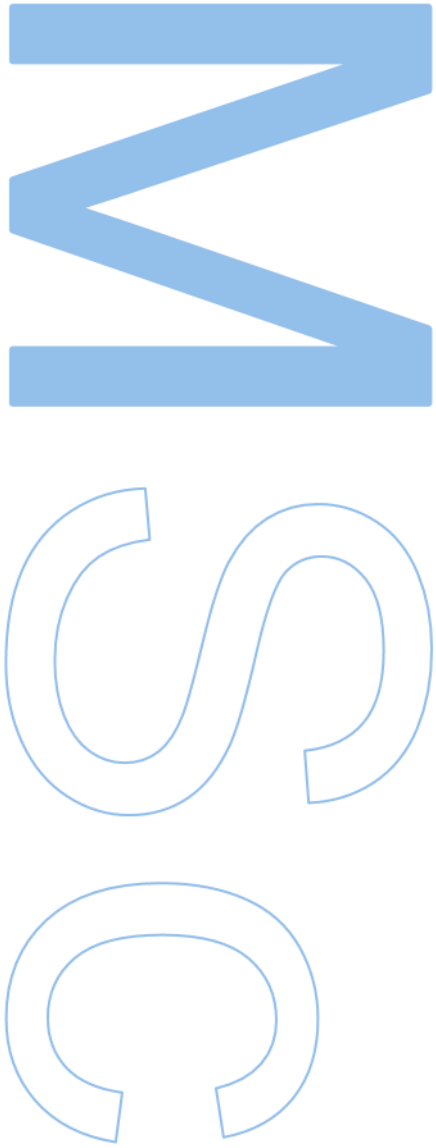
Dashboard NLP para análise de obras literárias



Arina Sanches
Mestrado em Ciência de Dados (Data Science)
Departamento de Ciência de Computadores
2023

Supervisor
Álvaro Figueira, Professor Auxiliar, Faculdade de Ciências

Dashboard NLP para análise de obras literárias

Arina Sanches

7 de dezembro de 2023

Abstract

Traditionally, literary analysis and comparison is done manually. It requires careful reading and deep understanding of the texts, and it can easily become a complex and time-consuming task for the reader. This type of task can greatly benefit from the use of Natural Language Processing (NLP) techniques. Over the past few decades, NLP has experienced exponential growth, leading to the development and widespread availability of public libraries that incorporate diverse Processamento de linguagem natural (NLP) techniques. Hence, an opportunity emerges to use these libraries for automating literary analysis.

We propose a system whose goal is not replace traditional reading, but to act as a supporting tool for literary analysis and comparison. The system explores how NLP techniques can be used to identify different elements of a novel. To promote a faster understanding of the novels, we use charts, networks and statistics to represent the analysis. This type of system can benefit a wide variety of user, but students, teacher and researchers of the field of literature are the one how could benefit the most.

Different libraries for NLP were studied and adapted to develop the proposed system. The sub-areas of NLP explored in this work include Named Entity Recognition (NER), topic modeling, keyword extraction, sentiment analysis, emotion detection, and text summarization. Through these approaches, we developed methods to perform various tasks such as extracting characters and their relationships, identifying points in the story with significant emotional variations, identifying relevant locations and dates in the stories, among others. Another important contribution of this work is the creation of a dashboard that presents these analyses to the reader in a simple and interactive manner through various charts.

We conducted a study of five classic novels, using the dashboard, to evaluate the performance of the system. The system achieved great results in character identification. It was able to identify all the main characters and their relationships in the history. We were also able to use the dashboard to identify the main locations of the story. Topic modeling and keyword extraction, were less successful, as the algorithms struggled to identify the main topics of the novels. The identification of temporal and date references encountered problems and did not achieve precise results. Onthe other hand, the dashboard panels dedicated to sentiment and emotion analysis were able to describe the emotional tone of the novel and correctely identified points of important sentiment variation.

Keywords - Data Science, Literary Analysis, Named Entity Recognition, Topic Modeling, Keyword Extraction, Sentiment Analysis, Emotion Detection, Text Summarization.

Resumo

Tradicionalmente, a análise e comparação de obras literárias é realizada de forma manual. Uma vez que exige uma leitura cuidadosa e uma compreensão profunda dos textos, pode facilmente tornar-se uma tarefa complexa e exigir muito tempo do leitor. Assim, conclui-se que este tipo de atividade muito pode beneficiar de técnicas de Processamento de linguagem natural (NLP). Ao longo das últimas décadas, a NLP vivenciou um crescimento exponencial, o que promoveu o seu desenvolvimento e disseminação através de bibliotecas públicas que oferecem suporte a aplicação das suas diversas técnicas. Surge assim a oportunidade de utilizar essas bibliotecas para a automatização da análise literária.

Neste trabalho, propomos um sistema, que tem como objetivo, não substituir a leitura tradicional e manual, mas sim ser uma ferramenta de apoio a análise e comparação de obras literárias. O sistema desenvolvido visa explorar como podemos utilizar técnicas de NLP para identificar elementos importantes das obras e representá-las através de gráficos, redes e estatísticas, a fim de promover um entendimento mais rápido das obras literárias. Este tipo de sistema pode ser útil para diversos tipos de leitores, porém oferece benefícios especialmente significativos para estudantes, professores e, de maneira geral, investigadores na área da literatura.

Para a criação do sistema, foi necessário a exploração e adaptação das técnicas de NLP, disponibilizadas pelas bibliotecas públicas, para o contexto aplicado da análise de obras literarias. As sub-areas da NLP, exploradas neste trabalho são: Named-entity recognition (NER), *topic modeling*, extração de *keywords*, análise de sentimentos, deteção de emoções, e sumarização de texto. Através destas abordagens, desenvolvemos métodos para realizar diferentes tarefas tais como, extrair personagens e as ligações entre eles, identificar os pontos da história com variações emocionais significativas, identificar os locais e datas relevantes nas historias, entre outros. Outra importante contribuição deste trabalho, é a criação de uma *dashboard*, que através de diferentes gráficos, apresenta ao leitor essas análises de uma forma simples e iterativa.

Para avaliar o sistema, foi feita uma análise do comportamento do sistema, em relação a cinco obras literárias clássicas. O sistema obteve resultados muito satisfatórios na identificação de personagens, tendo conseguido identificar todos os personagens principais e, para além disso, conseguiu identificar ligações entre personagens na trama. Os métodos aplicados também obtiveram sucesso na identificação de referências de localização. Os pontos menos bem sucedidos, temos a analise de tópicos e a extração de *keywords*, onde os algoritmos não conseguiram eficiente

identificar os principais tópicos e temas discutidos nas obras. A identificação de referencias temporais e de data, também enfrentaram problemas e não conseguiu alcançar resultados tão precisos. Por outro lado, houve sucesso do sistema na analise de sentimentos e emoções, através da qual é possível tanto compreender o tom emocional, assim como identificar pontos da história onde existem variações importantes no sentimento geral.

Palavras-Chave - Ciência de dados, Análise literária, Reconhecimento de Entidade Nomeada, Modelagem de tópicos, Extração de palavras-chave, Análise de sentimentos, Detecção de emoções, Sumarização de texto.

Agradecimentos

Agradeço ao meu orientador, Prof. Doutor Álvaro Figueira por ter me aceitado como orientanda, pelos valiosos conhecimentos transmitidos, por toda a dedicação e disponibilidade durante toda a elaboração da tese.

Agradeço aos meus pais, José Júlio Sanches e Alexandra Sanches, por todo o apoio e sacrifícios que fizeram para que eu pudesse chegar até aqui. Deixo também um especial obrigado, a minha irmã Eliane Swely Sanches, as minhas amigas Amanda Tavares e Lirielly Vitorugo, e ao meu amigo Carlos Silva, por todo o apoio durante a elaboração deste trabalho.

Por fim, gostaria de deixar um sincero obrigado a todos os amigos, familiares e professores que participaram e contribuíram para o meu percurso académico.

Dedico aos meus pais

Conteúdo

Lista de Figuras

Acrónimos

ATS sumarização automática de texto

CNN convolutional neural network

CRF campos aleatórios condicionais

CSV Comma-Separated Values

DNN deep neural networks

HMM modelos ocultos de Markov

IR Information Retrieval

k-NN k-nearest neighbors

LDA Latent Dirichlet Allocation

LOC location

NER Named-entity recognition

NE Named Entity

NLP Processamento de linguagem natural

NPMI normalized point-wise mutual information

ORG organization

PER person

PMI pointwise mutual information

RNN recurrent neural network

SVM support vector machines

TF-IDF Term Frequency - Inverse Document Frequency

VADER Valence Aware Dictionary and Sentiment Reasoner

Capítulo 1

Introdução

1.1 Contexto e motivação

Tradicionalmente, a análise literária é realizada manualmente, exigindo leitura atenta, anotações meticulosas e uma compreensão profunda dos textos. Tudo isso demora tempo e quanto maior for a obra mais tempo será necessário.

No entanto, com o avanço da tecnologia, o desenvolvimento do Processamento de linguagem natural (NLP), e a sua disseminação através de bibliotecas públicas em linguagens de programação, surgiu uma oportunidade para explorar formas automatizadas de promover essa análise.

Nesta tese pretendemos usar a tecnologia de NLP, disponível através dessas bibliotecas públicas e mais conhecidas, para automatizar a análise literária. O nosso sistema, embora não pretenda ser perfeito, representa uma primeira experiência na avaliação da capacidade da tecnologia comum de NLP para extrair aspetos que possam ser relevantes e auxiliem no entendimento mais rápido das obras literárias. O que pretendemos com este sistema é criar um ambiente de exploração em que características particulares são realçadas sob a forma de gráficos, redes e estatísticas de forma ao utilizador mais rapidamente os poder confrontar, quer entre capítulos da mesma obra, quer entre obras diferentes.

Assim, o nosso objetivo não é o substituir o prazer da leitura pela análise rápida e automática, mas sim fornecer uma ferramenta de análise a um tipo particular de leitor. Esse leitor é aquele que pretende: identificar aspetos genéricos de uma obra, comparar obras com estilos diferentes, identificar momentos de negativismo ou positividade ao longo da trama de uma obra, etc. Ou seja, o leitor que beneficiaria deste sistema de análise literária automatizada é multifacetado, abrangendo uma gama diversificada de pessoas com diferentes necessidades e objetivos.

Embora o sistema possa ser útil para uma ampla gama de utilizadores, os mais propensos a beneficiar são estudantes, professores e, duma forma geral, investigadores. Estes grupos têm uma necessidade intrínseca de análise literária comparada, e o sistema que propomos oferece uma solução para atender a essas necessidades. Não obstante, o sistema também pode ser

uma ferramenta valiosa para estudantes pré-universitários e entusiastas/amantes da literatura, proporcionando uma introdução acessível e rápida às obras literárias em análise. É importante novamente ressaltar que o nosso sistema não pretende substituir a análise literária tradicional, mas sim complementá-la, ou pelo menos servir de introdução. A interpretação e intuição humana, assim como a compreensão profunda, não podem, pelo menos atualmente, ser totalmente replicadas pelos algoritmos atuais. No entanto, ao combinar a rapidez de análise por NLP com a sensibilidade humana, podemos abrir novos horizontes na análise literária, tornando-a mais acessível, abrangente e enriquecedora.

1.2 Objetivos

O principal objetivo deste trabalho, é desenvolver um sistema que utilize técnicas de NLP, para fazer análise automática de diferentes elementos de obras literárias e apresentar ao utilizar representações visuais dos resultados das mesmas, de forma iterativa e de fácil compreensão. Para que isso seja possível, torna-se necessário:

- Utilizar técnicas de NLP no contexto aplicado;
- Perceber os problemas das bibliotecas disponíveis para aplicação destas técnicas e identificar alternativas para acomodar as deficiências encontradas;
- Utilizar as técnicas existentes como base para criação de algoritmos que permitam identificar aspetos mais particulares do contexto em estudo, como por exemplos a utilização do reconhecimento de entidades Named-entity recognition (NER), como base para descobrir a ligação entre personagens;
- Identificar personagens, ligações entre personagens e ações por eles praticadas;
- Identificar locais e datas relevantes para a história;
- Aplicar a análise de tópicos e *keywords*;
- Permitir a análise da variação do sentimento e emoções ao longo dos capítulos da história;
- Aplicar técnicas de sumarização de texto;
- Criar visualizações que melhor representem as análises realizadas;
- Organizar essas visualizações em uma *dashboard* interativa que permita ao utilizar fazer uma análise capítulo a capítulo das obras.

1.3 Estrutura e organização do documento

Este trabalho é composto por de seis capítulos, incluindo o capítulo atual, sendo divididos da seguinte forma:

- **No Capítulo 2**: apresentamos o estado da arte da área de NLP. Começando com uma introdução à NLP, as suas aplicações, técnicas de pré-processamento de texto e desafios. Em seguida, exploramos em várias subáreas da NLP, a modelagem de tópicos, detecção de *keywords*, NER, análise de sentimentos, deteção de emoções e sumarização de texto. O texto fornece uma compreensão geral dessas subáreas, suas aplicações e os desafios.

- **Capítulo 3**: descrevemos todo o *pipeline* que foi desenvolvido para possibilitar a análise das obras. Desde a preparação dos dados, até a aplicação das técnicas de NLP, para a geração de dados que serão posteriormente utilizados para alimentar uma *dashboard*. Neste capítulo, são apresentados detalhes das técnicas escolhidas durante o processo.

- **Capítulo 4**: o foco deste capítulo é a *dashboard*, desenvolvida para apresentar os resultados obtidos através da aplicação das técnicas de NLP. Detalhamos o funcionamento da *dashboard* e das visualizações que a compõem.

- **Capítulo 5**: neste capítulo, apresentamos as obras literárias que estão disponíveis para o utilizador do sistema. Apontando informações relevantes como o autor, género, ano de lançamento, uma breve sinopse do enredo, os principais temas abordados e o estilo narrativo de cada obra. Em seguida, exploramos o comportamento do sistema para a análise das obras, destacando pontos onde ele foi mais ou menos bem sucedido.

- **Capítulo 6**: por fim, no último capítulo apresentamos as nossas conclusões, as principais contribuições do trabalho, os desafios e as limitações encontradas, e possíveis trabalhos futuros.

Capítulo 2

Estado da Arte

2.1 Processamento de linguagem natural (NLP)

Diariamente são produzidos e coletados um grande volume de dados. No entanto, uma grande parte destes dados, são não estruturados. Realizar uma análise em grande escala, com o objetivo de obter *insights*, torna-se uma tarefa cada vez mais desafiadora. A NLP, é uma área de estudo que nasceu da necessidade de lidar com a vasta quantidade de dados em linguagem natural, isto é, escrita por seres humanos [7].

A NLP, é uma subárea da inteligência artificial, que utiliza de conhecimentos de ciência da computação, linguística e da matemática, para análise, representação e geração automática de dados em linguagem natural. Através de diferentes técnicas torna-se possível a tradução da linguagem natural para informação manipulável pelas máquinas [7, 13, 15]. De acordo com [7], podemos destacar sete tarefas que são de grande interesse para a área, sendo elas: a tradução, sumarização de texto, Information Retrieval (IR), extração de informações, *question answering*, *topic modeling* e *opnion mining*.

2.1.1 Componentes do processamento de linguagem natural

A NLP pode ser entendida como um processo dividido em diferentes componentes, estas componentes são ilustradas na figura 2.1. Neste contexto, a linguística é dividida em duas vertentes: a linguística computacional e a linguística teórica. A linguística computacional estuda e desenvolve algoritmos para análise e geração da linguagem natural. Em contrapartida, a linguística teórica tem como foco o desempenho linguístico e a competência gramatical [7].

A análise linguística é subdividida em: análise frásica e análise de estrutura do discurso e do diálogo. A análise frásica tem por objetivo definir o significado das frases, enquanto que análise de estrutura do discurso e do diálogo, procura compreender o significado do texto como um todo. Para compreender o significado de um texto, é preciso que para além de se compreender o significado de cada frase individual, se compreenda também as conexões existes entre as mesmas.

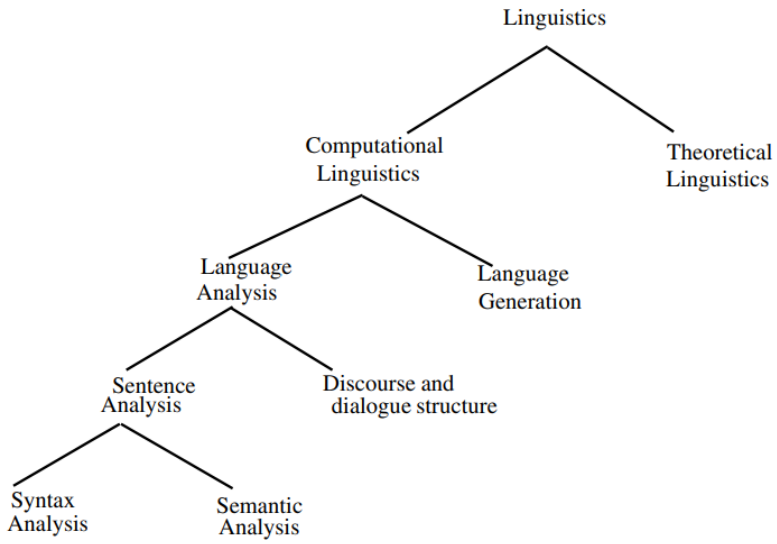


Figura 2.1: Adaptado de [7]

Muitas vezes essas conexões estão implícitas, o que dificulta a tarefa de identificá-las.

A análise de sintaxe e de semântica, formam a análise frásica. Sendo que a análise de sintaxe se concentra principalmente em duas tarefas:

- **Determinação da estrutura da frase**: preocupa-se com análise, interpretação e representação da estrutura gramatical das sentenças. Identifica os sujeitos, objetos, verbos e modificações presentes nas orações, assim como a função por eles desempenhada;
- **Regularização sintática**: refere-se ao processo de padronização ou normalização da estrutura sintática de frases ou textos. Isto pode ser alcançado de diferentes formas como a transformação de formas passivas em ativas, omissão de elementos da frase ou restauração de informações omitidas. A aplicação destas técnicas pode simplificar o processo de análise semântica.

2.1.2 Técnicas de Pré-processamento de dados

O pré-processamento dos dados é a primeira etapa num projeto de NLP. As técnicas aplicadas nesta etapa, têm por objetivo melhorar a qualidade dos dados e garantir que estes possam ser utilizados pelos algoritmos de NLP de forma eficiente [30].

2.1.2.1 *Tokenization*

É um processo através do qual uma sequência grande de caracteres é subdividida em grupos mais pequenos, geralmente separados por espaços em branco (formando assim as palavras) que possuem significado, e que são chamados de *tokens*.

Diferentes linguagens trazem diferentes desafios para este processo de divisão. Em línguas como o Português, em que as palavras são separadas por espaços, a divisão das frases em *tokens* torna-se mais simples. Em contrapartida, línguas como o chinês e o tailandês, são não assim segmentadas e as barreiras entre as palavras não são tão claras, e necessitam que técnicas mais sofisticadas sejam utilizadas [14].

2.1.2.2 Remoção de *stop words*

Num documento, conjuntos de palavras ocorrem em frequências diferentes. Algumas palavras ocorrem com grande frequência nos documentos, mas não são discriminantes para a compreensão do significado dos mesmos. Essas palavras são conhecidas como *stop words* e podem ser conectores, artigos, preposições, pronomes, entre outros. A remoção das *stop words* ajuda a diminuir a dimensionalidade do espaço de termos e com isso aumentar a eficiência dos algoritmos. Um dos desafios encontrados ao se realizar esta tarefa, é a criação da lista correta de *stop words* que devem ser removidas dos documentos [14, 30].

2.1.2.3 *Stemming* e *Lemmatization*

O processo de *stemming* procura reduzir as diferentes variantes das palavras ao seu radical. Isto acontece através da remoção dos afixos das palavras. Por exemplo, temos as palavras *connect*, *connected*, *connecting* do inglês, estas podem ser todas convertidas para *connect*. O *stemming* assume que as palavras são semanticamente relacionadas e, dependendo da linguagem em estudo, são realizadas diferentes análises morfológicas [30]. Em [14] somos alertados para dois principais erros que podem ocorrer durante a realização do *stemming*, sendo eles:

- *over stemming* : acontece quando duas palavras com diferentes radicais, são convertidas ao mesmo radical;
- *under stemming* : duas palavras que possuem o mesmo radical, são convertidas em radicais diferentes.

A *Lemmatization*, por outro lado, procura identificar o "lema"(a forma de dicionário) das palavras, por meio de análises morfológicas e do uso de vocabulários que querem um profundo conhecimento linguístico. Por exemplo, as palavra *"studies"* e *"studying"*, ao passarem pelo processo de *Lemmatization*, são convertidas para *"study"*. Enquanto que se fosse utilizado o *stemming*, as palavras seriam convertidas para *studi* e *study*, respectivamente [3, 17].

As técnicas de *stemming* e *lemmatization* ajudam a melhorar a eficiência dos algoritmos de NLP. O seu uso permite reduzir significativamente o tamanho do vocabulário e o número de índices, uma vez que um único índice pode ser utilizado para representar diferentes palavras que possuam o mesmo lema ou radical [3]. O processo de desenvolver uma abordagem para o *lemmatization*, é mais complexo e computacionalmente custoso do que o de desenvolver uma

abordagem de *stemming*. No entanto, as técnicas de *lemmatization* conseguem produzir resultados mais precisos [17].

2.1.3 Desafios e Tendências Futuras

A NLP é uma área que teve um crescimento exponencial nas últimas décadas, mas ainda enfrenta desafios significativos devido à natureza em constante evolução das línguas e suas complexas estruturas. Entre os desafios destacados por Khurana et al. [16] estão:

1. A ambiguidade é um desafio comum em NLP, pois a linguagem natural muitas vezes contém palavras ou frases com múltiplos significados. Os sistemas de NLP precisam utilizar contextos, regras gramaticais e técnicas de aprendizado de máquina para lidar com essa ambiguidade e determinar o significado correto com base no contexto da frase ou do texto;
2. Lidar com sinónimos e homónimos na linguagem também é um desafio, pois requer o desenvolvimento de algoritmos que considerem a diversidade de palavras usadas para expressar ideias semelhantes e a capacidade de interpretar corretamente o significado de palavras homónimas com base no contexto;
3. Identificar a presença de sarcasmo e ironia nas frases constituintes de um documento, é um desafio que ainda não foi completamente solucionado. Essas formas de comunicação muitas vezes envolvem significados não literais que podem ser difíceis de identificar de forma automática;
4. Os algoritmos de NLP também precisam estar preparados para lidar com componentes linguísticos que variam entre diferentes domínios, culturas e localizações geográficas. O significado ou existência de expressões ou palavras pode variar de acordo com a localização geográfica ou domínio onde são utilizadas.

Otter et al. [25] ressalta a necessidade de expandir a abrangência linguística nos métodos de NLP,uma vez que a maioria se foca na língua Inglesa e em menor escala no chinês mandarim. Em razão disso muitos aspetos linguísticos, que não estão presentes nas línguas acima mencionadas, não são explorados pelos sistemas de NLP atuais. Um dos fatores causadores desse problema, é o reduzido número de conjuntos de dados disponíveis para grande parte das línguas faladas no mundo, em razão disso também não é possível utilizar modelos de *deep learning* nesses casos.

2.2 Named-entity recognition (NER)

Nem todos os elementos de uma frase possuem a mesma relevância, alguns elementos carregam informações importantes, que são utilizados em muitas tarefas de NLP. Neste contexto, Named Entity (NE) são palavras ou frases que estão relacionados com um dado tópico e possuem um conjunto similar de atributos [18, 22].

li et al. [18] define NER como a identificação de NE em textos e a sua classificação em categorias predefinidas. NER pode ser utilizada para auxiliar diferentes aplicações da NLP, por exemplo *Question Answering*, tradução de texto, IR [22], *Semantic Annotation*, *Ontology Population* e *Opinion Mining* [20]. Atualmente os métodos desenvolvidos e as bibliotecas, permitem identificar uma variedade de NE, sendo que a maioria desses sistemas identifica as classes person (PER), location (LOC), organization (ORG), que dizem respeito a pessoas, locais e organizações, respetivamente [22].

2.2.1 Abordagens para NER

Ao longo dos anos, diferentes abordagens foram propostas para NER. Aqui, vamos abordar somente algumas das mais relevantes:

- **Rule-Based Named Entity Recognition**: Esta foi a primeira estratégia abordada para NER. Esses sistemas consistiam em um conjunto de regras de extração de entidades, *gazeteers* para várias categorias de entidades e um mecanismo de extração que aplicava essas regras e *lexicons* ao texto. As regras e *lexicons* eram criados manualmente ou a partir de alguns exemplos manuais. Embora sistemas baseados em regras fossem precisos, eles geralmente tinham uma cobertura limitada e funcionavam melhor em domínios específicos. A eficácia desses sistemas dependia da abrangência das regras e *lexicons*, e a incorporação de conhecimento mais profundo exigia esforço manual caro respetivamente [22];
- **Statistical Named Entity Recognition** : Através desta abordagem foi possível reduzir o esforço humano associado ao processo de NER e permitir maior flexibilidade na incorporação de aspetos linguísticos, como informações sintáticas. Para isso são precisos dois componentes principais, dados de treinamentos rotulados e um modelo estatístico, que é uma representação probabilística dos dados de treinamento. O modelo gerado é composto por parâmetros que mapeiam eventos linguísticos para probabilidades. Neste contexto diversas técnicas supervisionadas de *Machine learning* podem ser empregadas[18, 22];
- **Unsupervised Learning** As abordagens de aprendizagem não supervisionada, como o clustering, tem sido utilizadas para NER. Com esta abordagem é possível eliminar a necessidade de dados de treinamento rotulados. São usadas medidas de similaridade de contexto, recursos léxicos, padrões léxicos e estatísticas calculadas em grandes *corpora* para identificas as NER. Várias experiências demonstraram a eficácia e generalização dessas abordagens não supervisionadas em diversos domínios, como por exemplo na análise de texto biomédicos [18];
- **Deep learning**: Os métodos de *deep learning* têm-se destacado cada vez mais e alcançado resultados significativos entre as abordagens utilizadas no NER. O *deep learning*, uma técnica de *machine learning* que usa redes neuronais com várias camadas para aprender representações de dados em diferentes níveis de abstração [18];

- **Hybrid Systems**: são abordagens que combinam duas ou mais abordagens, a fim de oferecer maior flexibilidade e melhor desempenho dos modelos. Através desta abordagem, é possível adaptar-se às necessidades especificas de cada aplicação e domínio [22].

2.2.2 Desafios

Mohit and Imed [22] identificam a determinação das fronteiras para as NE e o reconhecimento das classes das NE, como os dois principais problemas enfrentados pelo NER. A ambiguidade inerente as linguagens dificultam a resolução destes problemas. Outro fator a ser levado em consideração, é a dependência em relação ao vocabulário e à especificidade do domínio em estudo. As línguas estão sempre em constante evolução e o conjunto de nomes próprios também. Para criar um sistema para NER confiável em um domínio especifico, são necessários dados rotulados e léxicos já definidos. No entanto, criar e manter esses recursos para várias áreas é caro e requer conhecimento linguístico e especialização no domínio.

2.3 Topic modeling

Um humano, consegue, com alguma facilidade, ler um dado documento e identificar quais os eventos ou conceitos que são nele abordados. No entanto, para uma máquina esta tarefa não é tão trivial. Também é preciso levar em consideração o volume de dados que temos a nossa disposição atualmente, em alguns cenários torna-se impraticável analisar e relacionais esses documentos [29]. O *topic modeling*, é uma subárea da NLP, em que as suas técnicas têm como objetivo identificar os tópicos abordados em documentos, podendo este ser emails, livros, publicações em redes sociais, artigos, ou outras formas de texto não estruturado. Neste contexto, um tópico é um conjunto de palavras que, conjuntamente, fazem referência a dado conceito [12].

As técnicas de *topic modeling* são aplicadas em diversas áreas, por exemplo na análise de artigo científicos, lidando com grandes coleções de documentos académicos [28]. Também são empregadas na análise de medias sociais, onde ajudam a compreender o comportamento dos utilizadores, a forma como interagem em comunidades e na extração de padrões em suas conversas e conteúdos compartilhados em plataformas digitais. Essas técnicas também encontram aplicação em campos como saúde, engenharia de software, geografia, ciências políticas, na linguística, entre outros [12].

A modelação de tópicos, apesar de sua popularidade, enfrenta desafios significativos, como problemas de otimização, sensibilidade ao ruído e instabilidade, que podem afetar a confiabilidade dos dados. A escolha adequada de métodos é crucial para extrair estatísticas significativas de um conjunto de dados. Embora as abordagens atuais tenham melhorado significativamente em relação a algoritmos anteriores, ainda requerem otimização e ajuste para resultados confiáveis, especialmente quando aplicadas a tipos específicos de dados e relações, como por exemplos, textos curtos, ou dados que são muito correlacionados, ou até com complexas relações estruturais [29].

2.3.1 Latent Dirichlet Allocation (LDA)

O LDA, é uma abordagem desenvolvida para *topic modeling* por Blei et al. [4]. Foi uma das primeiras técnicas utilizadas para *topic modeling* e continua até hoje a ser uma das mais utilizadas. Esta abordagem procura modelar documentos como sendo compostos por múltiplos tópicos, e onde cada tópico é definido como uma distribuição em um vocabulário fixo de termos [28]. Mais formalmente, podemos definir o LDA como um modelo probabilístico generativo, que representa documentos como misturas aleatórias de tópicos latentes, onde cada tópico é caracterizado por uma distribuição de palavras. O LDA utiliza a distribuições de Dirichlet para modelar a distribuição de palavras em tópicos e a distribuição de tópicos em documentos[12].

No entanto, o LDA enfrenta desafios que incluem:

- Esparsidade: esta questão é levantada quando o LDA é aplicado a *corpora* com um extenso vocabulário;
- Instabilidade: o algoritmo é muito sensível a mudanças no conjunto de dados ou na sua configuração;
- Inferência de parâmetros: é o processo de estimar os valores ótimos dos parâmetros do modelo.

O LDA assume alguns pressupostos que podem não ser válidos, o que pode comprometer o seu desempenho ao ser aplicado a conjuntos de dados reais. Uma das suposições é que o valor de "k"(número de tópicos) é fixo e conhecido. Na maioria das aplicações do mundo real, não existe um número ideal de tópicos previamente estabelecido. A determinação desse valor muitas vezes envolve a geração iterativa de modelos de tópicos com diferentes valores de "k"e a análise de métricas para comparação, um processo que muitas vezes é ineficiente e impreciso, especialmente em grandes conjuntos de dados. Outra suposição do LDA é que todos os tópicos são independentes entre si, o que não leva em consideração a correlação entre tópicos, algo comum em muitos tipos de dados, como texto e dados provenientes de redes sociais. Além disso, o LDA assume que documentos e palavras dentro dos documentos são independentes, o que pode não ser realista em muitos casos. [29]

2.3.2 Avaliação dos modelos em *topic modeling*

A avaliação da qualidade dos modelos identificados por *topic modeling* e a determinação do número ideal de tópicos representativos do texto são dois desafios para *topic modeling*. Ainda não existe um procedimento estatístico padrão para solucionar estes problemas [19]. No entanto, as métricas *perplexity* e *coherence* são frequentemente utilizadas para auxiliar na avaliação a qualidade dos modelos de tópicos gerados. A métrica *perplexity* é utilizada no treino dos modelos, para avaliar quão bem o modelo gerado, consegue descrever os documentos, utilizando os tópicos identificados. Quanto menor o valor de *perplexity*, melhor é considerado o modelo

gerado[9, 32]. Por outro lado, também é preciso garantir que os tópicos são semanticamente coesos e interpretáveis para os seres humanos. Para avaliar esses aspetos, utiliza-se a métrica *coherence*, que pode ser calculada com base no pointwise mutual information (PMI) [9] ou no normalized point-wise mutual information (NPMI) [23]. Quanto mais coerentes e relacionadas forem as palavras constituintes dos tópico, maior será o valor de *coherence*. Estudos têm demonstrado que a avaliação realizada através da *coherence*, em grande parte vai de encontro ao julgamento humano [9, 23].

2.4 Deteção de palavras-chave

Atualmente, produzimos e armazenamos uma vasta quantidade de documentos, acompanhada por um grande volume de informações associadas a eles. A análise de uma quantidade tão grande de documentos, torna-se difícil ou até mesmo impossível para um ser humano. Neste contexto, um conjunto de palavras que nos dão a indicação das principais características, temas e conceitos abordados em um documento, são uma valiosa ferramenta na realização desta tarefa [27].

As palavras-chave (*keywords*), que podem consistir em uma sequência de palavras ou em apenas uma única palavra, têm como objetivo representar o tema principal de um documento [26]. Em contrapartida, Siddiqi and Sharan [27] introduzem a distinção entre dois termos: *Keyphrase* e *Keyword*. De acordo com Siddiqi and Sharan [27], uma *Keyphrase* é uma composição de múltiplas palavras, enquanto uma *Keyword* é uma única palavra. É importante notar que a separação das palavras em uma *Keyphrase* pode levar a uma interpretação equivocada de seu significado, como no caso da expressão em inglês *'hot dog'*, onde essas duas palavras, quando combinadas, referem-se a um prato culinário, mas separadamente têm significados completamente diferentes.

A extração de *keywords*, aplicável em diversas áreas, como na leitura de artigos académicos, ajudam o leitor a ter identifica os principais temas abordados [27]. Na área de IR, são utilizados para definir consultas de sistemas [26]. Também podem ter um impacto positivo nas áreas de indexação automática, sumarização de texto, classificação de texto, *clustering* e filtragem [24].

A extração automática de *keywords*, pode ser feita através de diferentes abordagens, entre elas podemos destacar:

- **Baseadas em Regras Linguísticas**: Essa abordagem utiliza as características linguísticas das palavras e conduz análises léxicas, sintáticas, de discurso, entre outras. Embora capaz de alcançar resultados precisos, essa abordagem requer um profundo conhecimento de domínio e linguagem, assim como recursos computacionais [27].
- **Abordagens estatísticas**: Nesta abordagem, as *keyword*, são identificadas através de medidas estatísticas, como frequência da palavras ou estatísticas de n-gramas , e o Term Frequency - Inverse Document Frequency (TF-IDF) [24]. Entre as vantagens desta abordagem, podemos destacar, que ela pode ser independente de domínio [24] e são independentes do idioma [27]. Geralmente estes métodos não apresentam resultados tão

precisos quanto os métodos linguísticos, mas o grande volume de dados que temos a nossa disposição atualmente permite obtenção de resultados satisfatórios através desta técnica [27].

- **Abordagens de *machine learning***: Usualmente, essa abordagem se apoia em modelos de aprendizado supervisionado, os quais necessitam de um conjunto de dados de treino. Entretanto, em situações em que a obtenção ou criação desses conjuntos de dados é inviável, recorre-se a métodos de aprendizagem não supervisionada ou semi-supervisionada [27].
- **Abordagens Específicas do Domínio**: É possível também fazer uso dos conhecimento prévios de um domínio, como a sua ontologia, ou da estrutura do corpus em estudo, para identifica as *keywords*. Estas abordagens, são aplicadas a um corpus especifico de um dados domínio [27].

2.4.1 YAKE

Campos et al. [6] apresenta o YAKE!, uma abordagem utilizando aprendizagem não supervisionada, para a extração de *keywords*, em documentos não estruturados. O YAKE!, realiza a extração em documentos individuais, fazendo uso de *features* locais de texto e medidas estatísticas, como frequência e coocorrência de termos. O método desenvolvido pode ser aplicado a diferentes linguagens e domínios. Vale também ressaltar que, por utilizar documentos individuais, ele é independente da existência de um corpus. Atualmente, disponibilizam uma *demo*, uma API e um pacote *Python*, que possibilitam a utilização e/ou integração da abordagem desenvolvida pelo YAKE a outros projetos [5].

A abordagem desenvolvida para o YAKE!, é composta por cinco etapas [6]:

1. Pré-processamento de texto e identificação de termos candidatos;
2. Extração de *features*: os termos identificados na etapa anterior dão são representados por um conjunto de *features* estatisticas;
3. Cálculo da pontuação do termo: Utiliza-se de heurísticas para calcular um *score* que reflita a importância dos termos;
4. Geração de *keywords*, através de n-gramas, e cálculo do *score* das *keywords* candidatas, que reflitam a sua importância;
5. *Data deduplication* e criação de *ranking*: métodos de *Data deduplication* são utilizados para comparar a similariedade entre as *keywords*. As *keywords* são então ordenadas de acordo com os seus *scores*

2.5 Análise de sentimento

O crescimento e popularização da Internet, potencializou o crescimento das redes sociais, fórum de opinião e com isso a geração de um grande volume de dados sobre as opiniões dos utilizadores sobre diferentes temas. Paralelamente a isto, aconteceu um rápido crescimento da área de análise de sentimentos, tendo-se tornado uma das áreas mais estudadas da NLP. A análise de sentimentos tem também grande destaque nas áreas de *data mining*, *web mining* e IR. [8, 31].

A análise de sentimento é uma área de estudo que tem como objetivo identificar, extrair e quantificar informações subjetivas e emocionais sobre diferentes entidades, podendo estas entidade ser: produtos, organizações, serviços, textos, eventos, entre outros. Através da análise de sentimentos é possível definir se dado texto, apresenta opiniões positivas, negativas ou neutras. As técnicas da análise de sentimento, vêm sendo aplicadas em diferentes áreas, como: marketing, finanças, ciências políticas, comunicação, saúde e história [8, 31].

Segundo Zhang et al. [31] a análise de sentimentos pode ser feita em três níveis de granularidade:

- ***Document level*** : Aqui a unidade básica de informação é o documento como um todo, e o sentimento está associado integralmente a esse documento
- ***Sentence level***: Nesta vertente, o documento é separado em sentenças e a análise é realizada individualmente sobre cada uma das sentenças. É preciso também fazer a distinção entre frases que contêm opinião e as não opinativas. Apenas as frases classificadas como contendo opinião são consideradas na análise e atribuídas um sentimento.
- ***Aspect level***: Considerando, por exemplo, que um dado utilizador pode ter diferentes opiniões relativamente a diferentes aspetos de uma mesma entidade (podendo esta ser um produto, um serviço). A *Aspect level* procura atribuir sentimentos a diferentes aspetos das entidades em estudo. Para isso é preciso que primeiro se identifiquem as entidades e seus respetivos aspetos.

Dang et al. [8] apresenta três abordagens utilizadas para classificar o sentimento, baseadas em:

- ***Lexicon*** : As técnicas de análise de sentimento baseadas em léxicos foram pioneiras nesta área. Elas podem ser categorizadas em duas abordagens:
  - Baseadas em dicionários: utilizam dicionários de termos, que são disponibilizados em ferramentas como o WordNet;
  - Baseadas em *corpus*: Por outro lado, a análise baseada em corpora, utiliza análises estatísticas para a classificação do sentimento, removendo assim a necessidade da criação de um dicionário. Para isso ela faz uso de técnicas como k-nearest neighbors

(k-NN), campos aleatórios condicionais (CRF), modelos ocultos de Markov (HMM) , entre outras.

- ***Machine learning***: utiliza-se das técnicas de *machine learning* para realizar a classificação. Podendo-se dividir essas técnicas em dois grupos:
  - Modelos tradicionais: são as técnicas clássicas do *machine learning*, como por exemplo o *Naive Bayes*, o support vector machines (SVM). Esses algoritmos, fazem uso das de características lexicais, recursos baseados em léxicos de sentimentos, partes do discurso, adjetivos e advérbios, para realizar a tarefa de classificação.
  - Modelos de *deep learning*: estes métodos, tem mostrado muito promissores e conseguido alcançar resultados superiores aos dos modelos dos modelos tradicionais. Eles possibilitam a análise a nível do documento, da sentença e também do aspecto.Hoje, temos a capacidade de utilizar diversos modelos para essa tarefa, incluindo, por exemplo, o convolutional neural network (CNN), o deep neural networks (DNN), o recurrent neural network (RNN).
- ***Híbrido***: combina elementos das abordagens baseadas em léxicos com os da aprendizagem computacional.

2.5.1 Deteção de emoções

A deteção de emoções é uma subárea da análise de sentimentos que se concentra na identificação de diferentes emoções presentes em textos. Enquanto a análise de sentimentos procura classificar os sentimentos como positivos, negativos ou neutros, a deteção de emoções consegue alcançar um nível mais elevado de granularidade emocional. Tendo sido apresentado na literatura a existência de oito emoções básicas e que são alegria, tristeza, raiva, medo, confiança, nojo, surpresa e antecipação. Geralmente, para realizar a deteção de emoções, é adotada uma abordagem baseada em léxicos, embora abordagens baseadas em *machine learning* também sejam empregadas [21].

Acheampong et al. [1] destaca a importância das emoções nas experiências pessoais das pessoas e em suas interações com o ambiente, especialmente no contexto dos negócios e do comércio eletrónico. É enfatizada a necessidade de identificar as emoções expressas pelas pessoas, para fornecer recomendações personalizadas e melhorar a satisfação do cliente. Além disso, menciona a aplicação da deteção de emoções em outras áreas, como na saúde por exemplo, onde foi utilizada a deteção de emoções para analisar notas suicidas.

A deteção de emoções em texto ainda está em um estágio menos desenvolvido em comparação com a deteção de emoções em áudio, imagens e outros métodos multimodais. Isso ocorre porque os textos podem não apresentar pistas claras das emoções, ao contrário de outros métodos. Além disso, identificar emoções em textos curtos, *emojis* e textos com erros gramaticais é um desafio, sendo também importante considerar a evolução contínua das línguas. As técnicas disponíveis para essa tarefa, assim como os dicionários de emoções, ainda precisam amadurecer significativamente.

2.6 Sumarização de texto

A sumarização automática de texto (ATS) é a tarefa de criar um resumo conciso e fluente de um ou mais textos, preservando o conteúdo e o significado essenciais, enquanto reduz significativamente o tamanho e as repetições do texto original. A ATS é aplicável a diversos tipos de texto, incluindo documentos escritos, áudio e hipertexto [2, 10].

2.6.1 Classificações para sumarização de texto

El-Kassas et al. [10], apresenta diferentes maneiras de classificar sistemas de ATS, com base em diferentes aspetos:

- **Classificação baseada na estratégia de sumarização**: Existem três abordagens principais para a sumarização automática de texto: extrativa, abstrativa e híbrida. A abordagem extrativa seleciona secções importantes do texto e gera um resumo preservando as sentenças originais, dependendo apenas da extração de sentenças do texto original. Em contraste, as abordagens de sumarização abstrativa geram um resumo que não apenas extraia frases do texto de origem, mas também as reformule para criar um resumo mais conciso e fluente, interpretando e examinando o texto usando técnicas avançadas de processamento de linguagem natural para gerar um novo texto mais curto que transmita as informações críticas do texto original. Embora os resumos criados por humanos geralmente não sejam extrativos, a maioria das pesquisas em sumarização hoje se concentra na sumarização extrativa, pois os resumos puramente extrativos geralmente produzem melhores resultados em comparação com os resumos abstrativos automáticos, devido às complexidades associadas à sumarização abstrativa. De fato, muitos sistemas abstrativos ainda dependem de um componente de pré-processamento extrativo para produzir o resumo final. Além disso, existe a abordagem híbrida que combina elementos das abordagens extrativa e abstrativa, buscando combinar as vantagens de ambas [2, 10, 11];
- **Classificação baseada no tamanho dos dados de entrada:** A sumarização de texto pode ser realizada a partir de um único documento ou de múltiplos documentos. No entanto, quando se trata da sumarização de múltiplos documentos, surgem desafios adicionais, como a gestão de redundâncias, a abrangência da informação, considerações temporais, a definição da taxa de compressão, entre outros [10];
- **Classificação baseada no tipo de documento gerado:** pode ser Genérica ou *Query-base*. A abordagem genérica gera um sumario com o objetivo de extrair as informações mais relevantes, uma noção geral de um ou mais documentos. Por outro lado, a sumarização *Query-base* tem por objetivo extrair informações relacionadas com um tópico de pesquisa. Através da abordagem é possível extrair informação de um grupo de documentos homogéneos, retirados de um grande corpus como resultado de uma consulta [10, 11];

- **Classificação baseada na linguagem de sumarização:** de acordo com este critério, a sumarização pode ser, *Monolingual*, *Multilingual* ou *Cross-Lingual*. Quando a linguagem do texto original e do resumo são iguais, o sistema de sumarização é considerado *Monolingual*. Quando o texto original e o texto sumarizado, são escritos em múltiplas línguas a sumarização é *Multilingual*. Por fim, quando o texto original é escrito em uma língua, mas o texto sumarizado é escrito em outra, o sistema de sumarização é *Cross-Lingual* [10, 11];

- **Classificação baseada no algoritmo de sumarização:** os métodos de sumarização podem ser supervisionados ou não supervisionados. Os algoritmos supervisionados requerem a existem de um conjunto de dados de treino rotulados, enquanto que os métodos não supervisionados utilizam outras técnicas, como a utilização de heurísticas e clusterização para eliminar a necessidade de dados de treinamento rotulados [10, 11];

- **Classificação baseada no conteúdo do resumo:** pode ser indicativa ou informativa. Sumarizações indicativas oferecem uma visão geral do conteúdo, destacando tópicos gerais, enquanto sumarizações informativas fornecem informações mais detalhadas do documento original. A escolha entre esses tipos depende dos objetivos e necessidades do utilizados, as indicativas são úteis para obtenção de uma visão geral do tema abordado nos documentos e as informativas fornecem um compreensão mais profunda do conteúdo [10, 11];

- **Classificação baseada no tipo do resumo:** pode ser do tipo *Headline*, *Sentence-Level*, *Highlights*, ou *Full Summary*. Nesta abordagem o tamanho dos resumos torna-se progressimante maiores e fornecendo informações mais detalhadas. *Headline* produz um resumo extremamente conciso, que geralmente consiste em uma frase curta, já o *Sentence-Level* geralmente é constituído por uma única frase, fornecendo as informações chaves dos temas abordados. Os resumos *Highlights*, são um pouco mais longos do que os de nível de sentença e geralmente incluem várias sentenças, pequenos parágrafos ou texto na forma de *bullet points*. Por fim, os resumos *Full Summary* são mais detalhados e abrangentes. Eles abordam as principais ideias, detalhes e nuances do texto original, muitas vezes replicando o conteúdo original em um formato mais conciso [10].

- **Classificação baseada no domínio:** podem ser Gerais ou específicos de domínio, no primeiro o sistema de ATS sumariza documentos de diferentes domínios, enquanto que no segundo, são sumarizados documentos de um único domínio [10].

2.6.2 Desafios

A ATS enfrenta ainda muitos desafios, como a sumarização de múltiplos documentos ou documentos que sejam muitos longo (como livros), onde se enfrentam desafios como redundância, dimensões temporais, coreferências, entre outros. Decidir quando parar ou continuar um resumo, não é algo que seja facilmente definido, é preciso que se aprimorem técnicas para auxiliar nesta tarefa. Importante também ressaltar a necessidade dos sistemas de ATS fornecerem suporte a um maior leque de línguas, atualmente a maioria dos sistemas tem maior foco na língua Inglesa.

Outro grande desafio é a avaliação dos resumos gerados, a definição da qualidade de um resumo é uma tarefa subjetiva, até mesmo para os seres humanos. Pesquisadores continuam estudando novas técnicas para aperfeiçoar e otimizar a avaliação dos sistemas propostos. Por fim, estudos mostram que assim como outras áreas ligadas a NLP, a sumarização de texto, pode beneficiar do uso de métodos de *deep learning*, mas nem sempre encontram-se disponíveis grandes conjuntos dados de treino que estes métodos requerem [10].

Capítulo 3

Bases para o Desenvolvimento

Neste capítulo, exploramos como podemos utilizar diferentes abordagens da ciência de dados para a análise de obras literárias. A figura 3.1, ilustra o fluxo de processamento do sistema proposto por este trabalho. Começando pela leitura dos dados, o sistema esta preparado para receber arquivos no formato "txt"e é necessário que a divisão dos capítulos seja sinalizada com a palavra "Chapter"seguido do número do capítulo. Após a leitura dos dados eles passam pela etapa de preparação de dados. Posteriormente, aplicamos modelos de *Named-entity recognition*, *Topic modeling*, Deteção de *Keywords*, análise de sentimentos e deteção de emoções. Esses processamentos são feitos de forma automática , através de *scripts* e não requerem intervenção humana. Por meio dessas abordagens, procuramos identificar os personagens e as suas interações, os locais o tempo em que se passa a narrativa, e sentimentos e emoções transmitidos. Todos os códigos foram implementados utilizando a linguagem de programação *Python*, e os dados resultantes das abordagens desenvolvidas são armazenados em arquivos no formato Comma-Separated Values (CSV). Os dados gerados são posteriormente utilizados para a criação de gráficos, que compõe uma *dashboard*. A *dashboard* desenvolvida, tem por objetivo auxiliar o utilizar na compreensão das análises das diferentes obras literárias.

3.1 Preparação dos dados

O pré-processamento dos dados, é a primeira etapa das diferentes análises que foram realizadas durante este projeto. Sendo de suma importância, uma vez que são realizadas um conjunto de atividades, que permitem melhorar a qualidade dos dados e também transformá-los de forma a ser possível utilizá-los em processamentos futuros.

Um *pipeline* de pré-processamento de dados para Processamento de linguagem natural (NLP), pode ser composto por diferentes etapas. Nós começamos pela segmentação dos dados. É necessário adaptar a forma como as obras são apresentadas aos algoritmos, de forma a otimizar os resultados. Em razão disso, neste trabalho uma obra pode ser analisada tanto capítulo a capítulo, como frase a frase, dependendo do contexto. Através de diferentes testes, pudemos

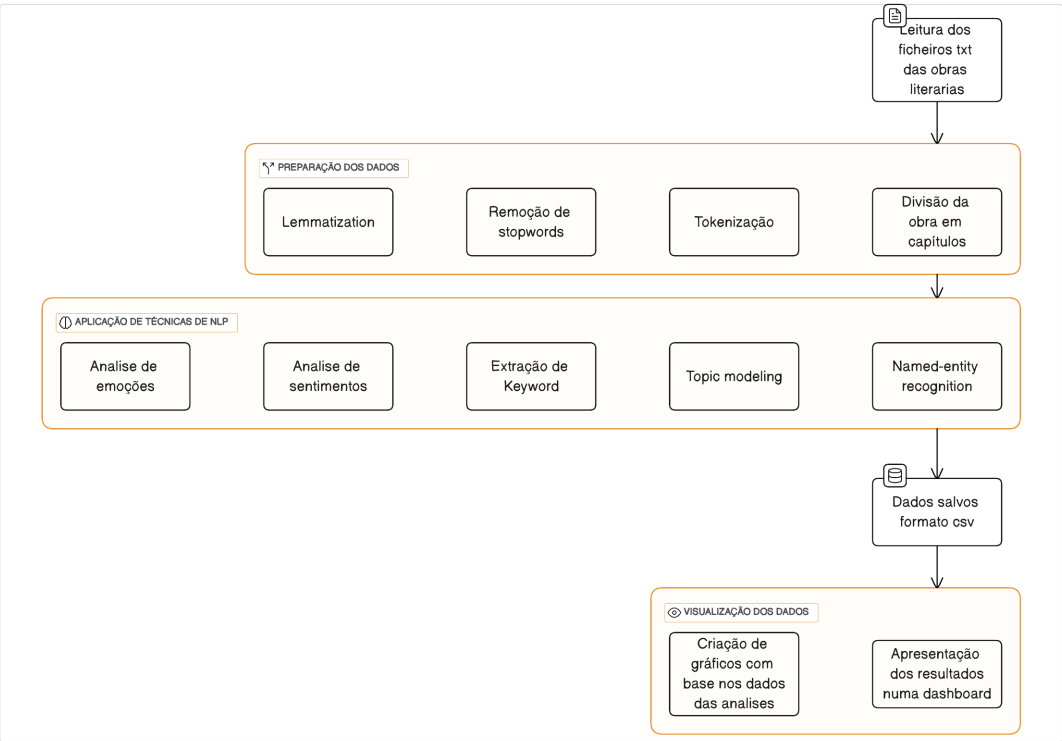


Figura 3.1: Fluxo de processamento do sistema proposto

constatar que dividir as obras em porções menores prejudicava o resultado final.

Outras etapas que foram realizadas incluíram a *tokenização* das frases, onde estas foram divididas em palavras, a remoção de palavras com menos de três letras e das *Stopwords*, que são palavras que ocorrem com muita frequência no texto, mas não são relevantes para a da análise dos documentos, e por a aplicação de técnicas *lemmatization*, que reduzem as palavras ao seu lema.

3.2 Named-entity recognition (NER)

O reconhecimento de entidades, foi um dos pontos-chave deste projeto, uma vez que foi a base para a criação de outros algoritmos que permitem relacionar diferentes aspetos das histórias. Existe diferentes tipos de entidades que podem ser identificadas num documento. Por exemplo, na frase "Em 1994, Jeff Bezos fundou a Amazon, o gigante do comércio eletrónico, na garagem de sua casa em Seattle", "Jeff Bezos"é uma entidade do tipo pessoa, "1994"uma entidade do tipo data, "amazon"e "Seattle"são entidades do tipo organização e local repetivamente.

A biblioteca *spacy* do python, foi utilizada para o NER. O *spacy* disponibiliza um modelo convolutional neural network (CNN), projetada para a língua inglesa, treinado com o OntoNotes, um grande corpus de texto anotado com várias informações linguísticas, incluindo sobre as Named Entity (NE).Este modelo é capaz de reconhecer diversos tipos de entidades, no entanto, para

este projeto, estamos considerando apenas um subconjunto delas:

- *DATE*: datas absolutas ou relativas e periodos;
- *EVENT*: eventos desportivos, guerras, desastres naturais, etc;
- *FAC*: prédios, aeroportos, autoestradas, pontes, etc;
- *GPE*: países, cidades, estados;
- *LOC*: localizações não relacionadas a entidades geopolíticas, montanhas e corpos de água;
- *MONEY*: valores monetários, moedas;
- *ORDINAL*: número ordinal;
- *ORG*: empresas, agências, instituições, etc;
- *PERSON*: pessoas;
- *QUANTITY*: medidas de peso ou distância;
- *TIME*: horas, referências temporais.

Neste projeto, optamos por utilizar um nível de granularidade mais baixo do que as classes de entidades disponíveis no *spacy*, o que nos levou a agrupar algumas delas. Por exemplo, as entidades *FAC*, *GPE* e *LOC* foram unificadas sob a categoria *PLACE*, que representa as localizações identificadas nas obras literárias. As categorias *MONEY*, *ORDINAL* e *QUANTITY* foram agrupadas como números, abrangendo diferentes tipos de números. Quanto às entidades *DATE* e *TIME*, elas foram utilizadas tanto de forma individual quanto agrupadas sob a categoria tempo e data. Já as entidades das classes *PERSON*, *EVENT* e *ORG* foram utilizadas individualmente.

3.2.1 Datas, Marcos temporais e Locais

As entidades temporais, de data e locais ajudam o utilizador do sistema a conhecer o contexto onde se passa a história e assim compreender melhor a mensagem.

Inicialmente, identificámos todas as datas, marcos temporais e locais, presentes em cada um dos capítulos, juntamente com a frequência com a qual foram mencionadas. Posteriormente utilizamos essas frequências como indicativos da relevância destas entidades para a obra.

3.2.2 Personagens

As entidades do tipo *PERSON*, que fazem referencia a pessoas, foram utilizadas para identificar os personagens pertencentes as obras literárias. Tendo também sido utilizadas para criar redes de relacionamentos entre os personagens e a identificar as ações por eles realizadas.

3.2.2.1 Rede de relacionamento entre personagens

Para criação da rede de relacionamentos, testámos diferentes abordagens, num primeiro momento tentámos fazer um algoritmo que define as ligações, utilizando o capítulo completo. Porém, notamos que esta abordagem, produzia resultados muito generalistas, não permitindo conhecer melhor o relacionamento entre os diferentes personagens. Por esta razão, seguimos com a analise frase à frase, onde conseguimos captar um maior número de detalhes. No algoritmo desenvolvido, consideramos que dois personagens interagem caso estejam presentes na mesma frase. Sendo assim, cada personagem é representado por um nó do gráfico, onde o tamanho desse nó é proporcional ao número de vezes em que ele é referido na obra. Cada interação entre personagens é representada por uma aresta e o peso é definido de acordo com o número de vezes que os personagens interagem. Os valores referentes ao peso da aresta e ao tamanho do nó, são normalizados pelo valor máximo do peso da aresta.

3.2.2.2 Ações dos personagens

Para compreender uma história, é fundamental obter um conhecimento mais aprofundado sobre os personagens e os eventos que se desenrolam ao longo da narrativa. Nesse sentido, desenvolvemos um algoritmo com o propósito de estabelecer conexões entre os personagens, suas ações distintas, datas relevantes e os respetivos locais em que ocorrem.

Para identificar as ações, dividimos as obras em capítulos e em seguida em frases. Através da biblioteca *spacy*, identificamos os sujeitos das frases e os verbos a eles associados. Caso os sujeitos identificados fossem correspondentes a alguma das entidades do tipo pessoa, que já tinham sido previamente identificadas, estes eram considerados personagens, e os verbos das respetivas frases, as ações praticadas por estes personagens, neste caso o "evento". Consideramos igualmente relevante associar os eventos ás datas e locais. Para isso revisamos as frases anteriores àquela em que o evento foi identificado, partido da frase mais próxima e continuando para as mais distantes, até encontrarmos a primeira referência temporal e espacial, que tivesse sido previamente identificado pelo NER, como sendo uma entidade do tipo temporal ou de local.

3.3 Topic modeling

Aplicamos uma técnica de topic modeling pois a extração de tópicos e temas relevantes presentes na obra fornece uma visão resumida e estruturada dos principais assuntos abordados pelo autor. O objetivo é ajudar a descobrir padrões ocultos e tendências ao longo do texto, facilitando a identificação de elementos como enredo, cenários e questões temáticas recorrentes.

A análise dos tópicos foi feita capítulo à capítulo. Utilizámos um modelo pré-treinado, da biblioteca *gensim*, que tem por base o Latent Dirichlet Allocation (LDA). Porém, antes de ser possível utilizar o modelo, foi preciso definir o número ideal de tópicos para cada um dos

capítulos. O processo de encontrar o número ideal de tópicos é crucial para obter resultados significativos e representativos, é essencial determinar o número correto de tópicos que melhor descrevem a estrutura subjacente dos documentos. Para definir o número de tópicos, utilizamos de um processo iterativo e automatizado, onde para cada um dos capitulo, calculamos o valor da métrica *coehrence*, para modelos criados com o número de tópicos entre dois e dez. Em seguida, estes valores de *coehrence*, foram utilizamos por uma função que faz uma aproximação do método do cotovelo, para definir o número ideal de tópicos a ser utilizado em cada um dos capítulos. Na figura 3.2, apresentamos alguns exemplos das escolhas feitas pelo algoritmo de decisão do número de tópicos.

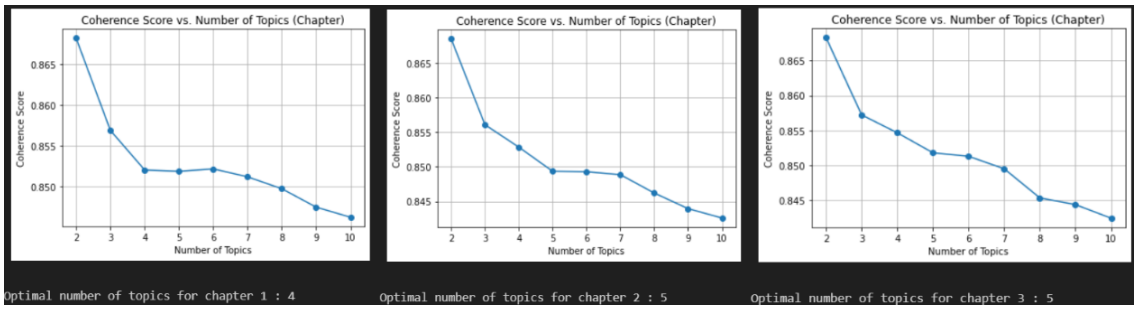


Figura 3.2: Gráfico dos valores de *coherence* para diferentes números de tópicos e capítulos

3.4 Deteção de Keywords

Realizamos a identificação das principais palavras-chave de cada capítulo. As palavras-chave representam conceitos importantes e relevantes presentes nas obras. A sua deteção auxilia o utilizar a construir uma visão abrangente e rápida sobre o conteúdo e os temas abordados na obra literária.

Utilizamos a biblioteca *YAKE* que disponibiliza um modelo não supervisionado de extração automática de palavras-chave. Esse modelo utiliza características estatísticas do texto, extraídas de documentos individuais, para selecionar as palavras-chave mais importantes de um texto.

Testámos diferentes números de palavras para as *keywords* e obtivemos os melhores resultados com *unigrams* e *bigrams*. Utilizando n-grams superiores a estes, as keyword se tornavam cada vez menos indicativas ou informativas em relação aos temas abordados nos capítulos.

3.5 Análise de sentimentos e emoções

Através destas duas análises, é possível compreender nuances emocionais que permeiam a narrativa e a construção dos personagens, enriquecendo a experiência do leitor e proporcionando uma compreensão mais completa da mensagem transmitida pelo autor. Ao analisar a polaridade dos sentimentos presentes na obra, é possível uma maior compreensão sobre o tom emocional

da história. Além disso, técnicas mais avançadas podem classificar emoções específicas, como alegria, tristeza, raiva, medo, entre outras, permitindo uma análise mais detalhada das emoções envolvidas.

3.5.1 Análise de sentimentos

Para a análise de sentimentos utilizámos a biblioteca a Valence Aware Dictionary and Sentiment Reasoner (VADER), uma ferramenta de análise de sentimento baseada em léxicos e regras, especialmente adaptada para identificar sentimentos. O VADER classifica as emoções utilizando 4 pontuações: uma para o sentimento positivo, outra para o sentimento negativo, uma para o sentimento neutro e, por fim, o *compound*. O *compound* é uma métrica que calcula a soma de todas as avaliações lexicais, normalizadas entre -1 (mais extremamente negativo) e +1 (mais extremamente positivo).

Criámos um algoritmo que analisa os sentimentos frase a frase. O VADER é usado para calcular o *compound*. Este valor é então acumulado pelo algoritmo ao longo das frases e dos capítulos. Escolhemos esta abordagem pois permite identificar momentos de intensidade emocional, momentos de calmaria ou estabilidade emocional e até mesmo mudanças significativas na atmosfera emocional da narrativa. Os pontos de subida podem estar associados a cenas ou eventos emocionalmente carregados, como clímax, revelações emocionais ou momentos de felicidade intensa. Por outro lado, os momentos de estabilidade podem representar momentos de transição ou desenvolvimento dos personagens, onde a história está mais focada na construção dos cenários e do enredo, sem grandes oscilações emocionais. Já os pontos de queda podem indicar momentos de tristeza, conflito ou desespero na história, onde os personagens enfrentam desafios significativos ou passam por momentos de perda e desilusão. Além disso, ao identificar esses pontos de subida, estabilidade e queda dos sentimentos, é possível observar a estrutura emocional da obra literária como um todo.

3.5.2 Deteção de emoções

A analise de emoções, foi feita utilizado a biblioteca *nrclex*. O NRCLexicon é um projeto pypi aprovado pelo MIT, que prevê os sentimentos e emoções de um determinado texto. Esta biblioteca abrange cerca de 27.000 palavras e fundamenta-se no léxico afetivo do National Research Council Canada (NRC), combinado com os conjuntos de sinónimos da biblioteca NLTK WordNet. Basicamente, consegue detetar as seguintes emoções a raiva, a antecipação, a confiança, a surpresa, a tristeza, o nojo, a alegria e o sentimento positivo e negativo. Além disso, também atribui uma percentagem de acordo com a intensidade de cada uma das emoções. Realçamos que a análise das emoções é feita capítulo a capítulo.

3.6 Sumarização de texto

A utilização de técnicas de sumarização de texto teve como principal objetivo apresentar ao utilizador uma síntese concisa, e informativa de cada um dos capítulos disponíveis na *dashboard* e com isso auxiliar na compreensão dos temas e desenvolvimentos da trama.

Para realizar esta tarefa, utilizamos a biblioteca *Sumy*, que disponibiliza um algoritmo de sumarização baseado em grafos conhecido como *TextRank*. Este algoritmo é inspirado no *PageRank* do Google e é usado para determinar a importância relativa das palavras e frases em um texto.

Capítulo 4

A interface do sistema

Neste capítulo, apresentaremos em detalhes a *dashboard* desenvolvida durante este projeto. O principal objetivo da *dashboard* é apresentar representações visuais dos resultados das análises discutidas no Capítulo 3, tornando as informações mais acessíveis e compreensíveis para os usuários. A *dashboard* foi desenvolvido utilizando a biblioteca R e o pacote *Shiny*, nela o utilizador pode explorar diferentes aspetos das obras.

A figura 4.1 é um esquema alto nível do funcionamento da *dashboard* desenvolvida. Através de um menu, localizado na barra lateral esquerda, o utilizador pode escolher entre diferentes obras previamente disponibilizadas na *dashboard*, no entanto só é possível visualizar uma obra e um capítulo por vez. Nesse sentido, é obrigatório que o utilizador selecione uma obra e, em seguida, escolha um capítulo ao utilizar a *dashboard*. Algumas visualizações possuem *tabs*, que permitam ao utilizador trocar a perspetiva da análise, podendo ver a informação relativa a um único capítulo, ou relativo a todos os capítulos da obra.

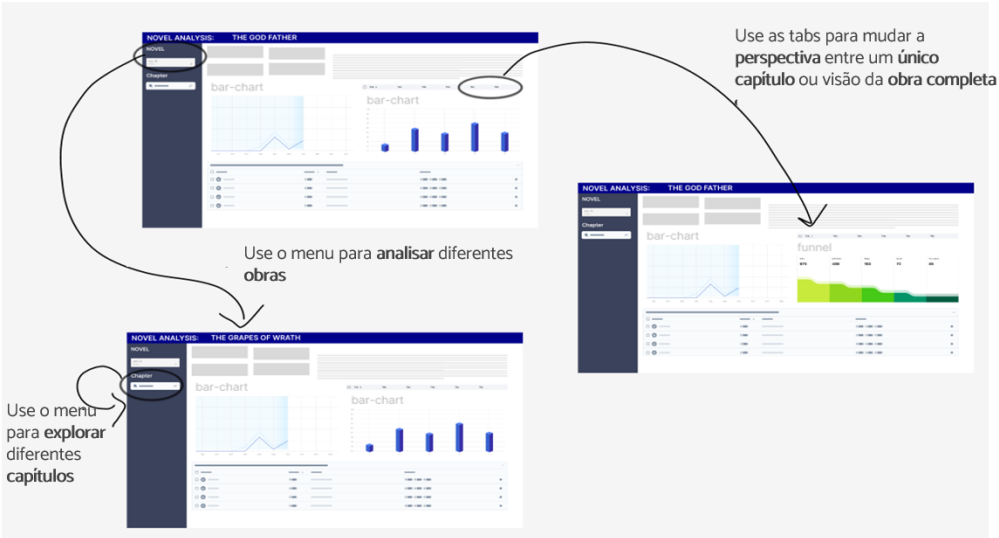


Figura 4.1: Wireframe do funcionamento da dashboard

Na figura 4.2 apresentamos um recorte da área superior da *dashboard*. Com o objetivo de

auxiliar na contextualização do utilizador, fornecemos algumas informações na barra lateral, incluindo o género da obra, o nome do autor, o ano e o pais de publicação, assim como a contagem do número total de palavras da obra e a contagem do número de palavras em cada capítulo. Para além disso, para cada um dos capítulos selecionados, mostramos um resumo correspondente.

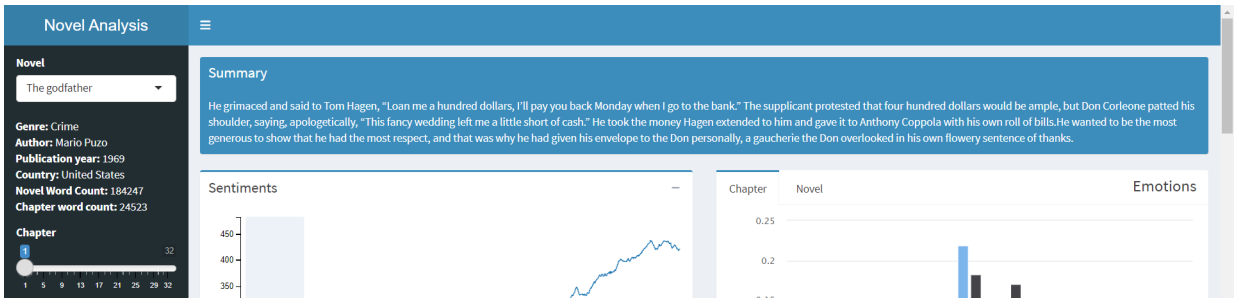


Figura 4.2: Recorte da área superior da *dashboard*

4.1 Personagens

Os personagens são as entidades que movem a história, através deles os temas, conflitos e mensagens do autor são explorados e comunicados. Em razão disso é dado a eles um grande destaque na *dashboard*. As personagens apresentadas aqui, foram identificados pelos algoritmos desenvolvidos com base em algoritmos de reconhecimento de identidades. Nos gráficos relativos aos personagens, são apresentados apenas os personagens que foram identificados em pelos menos 3 capítulos.

Na figura 4.3 foi utilizado o *TreeMap*. Os personagens que foram identificados no capítulo selecionado pelo utilizador, são destacados em azul, já os outros personagens aparecem em quadros cinza. Abaixo dos nomes de cada personagens, aparece a contabilização do número de capítulos em que o personagem foi identificado.

Ao longo da história, os personagens interagem entre si em maior ou menor escala. Um personagem interagir com um grande numero de personagens, é um indicador de que ele seja um dos personagens centrais e de grande relevância da trama. Na *dashboard*, utilizamos dois gráficos de redes distintos para representar as interações entre personagens.

A figura 4.4 e a figura 4.5 apresentam as duas visões do gráfico de interações entre personagens disponíveis ao utilizador na *dashboard*. O utilizador pode escolher entre duas visões para analisar as interações entre personagens, através do menu de navegação presente no canto superior esquerdo da figura 4.4. Escolhendo a opção "*Chapter*", lhe são apresentadas apenas as interações existentes no capítulo atualmente selecionado, isso é ilustrado figura 4.4. Por outro lado, caso ele escolha a opção "*Novel*", ele consegue visualizar todas as interações entre personagens presentes na obra. A figura 4.5 apresenta esta opção. Nos gráficos, cada nó representa um personagem, sendo o tamanho do nó proporcional ao número de menções ao personagem. As arestas representam as interações entre personagens, a espessura da linha é em função do número de interações entre

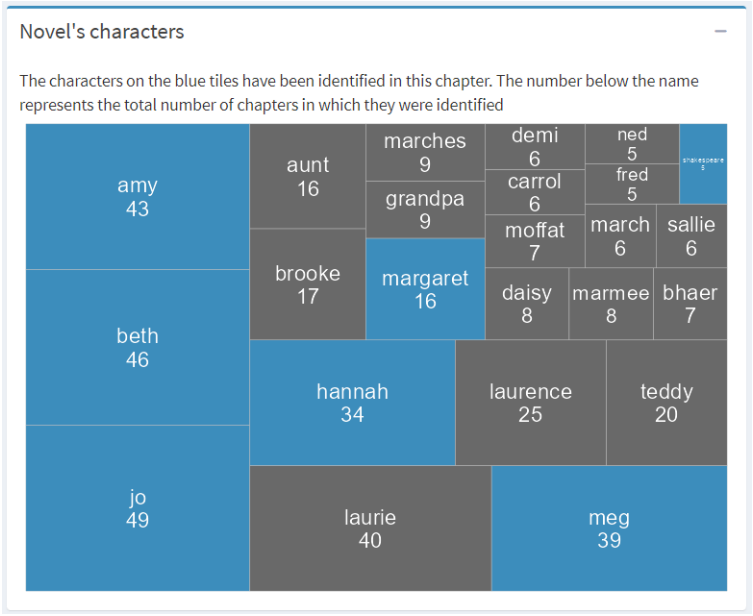


Figura 4.3: Gráfico dos personagens mais relevantes da obra literária

personagens, aumentando com o número de interações.

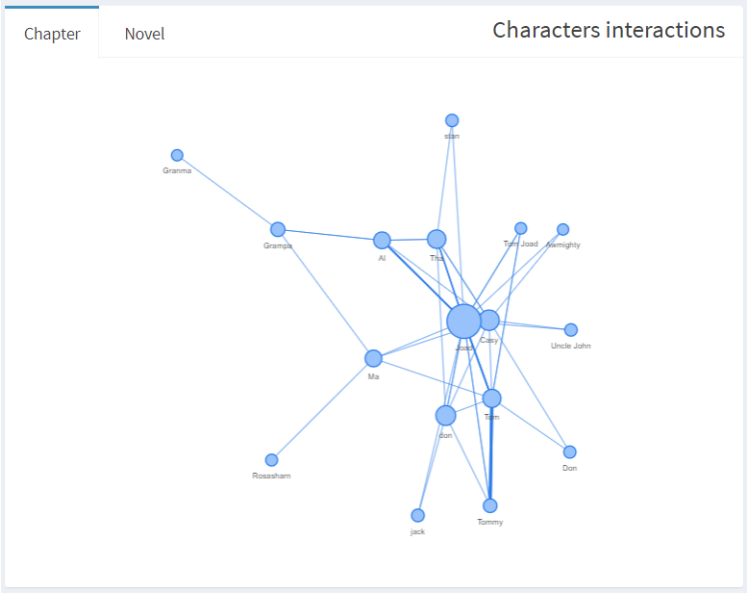


Figura 4.4: Rede de relacionamento entre os personagens no capítulo em análise

Exploramos também as ações praticadas pelos diferentes personagens. Ao escolher um capítulo, a tabela e a lista de personagens da figura 4.6 são atualizadas. Na lista de personagens, constam os personagens que o algoritmo conseguiu identificar as suas ações. Na tabela, são apresentadas as ações, os lugares onde estas decorreram, a data e a referência temporal.

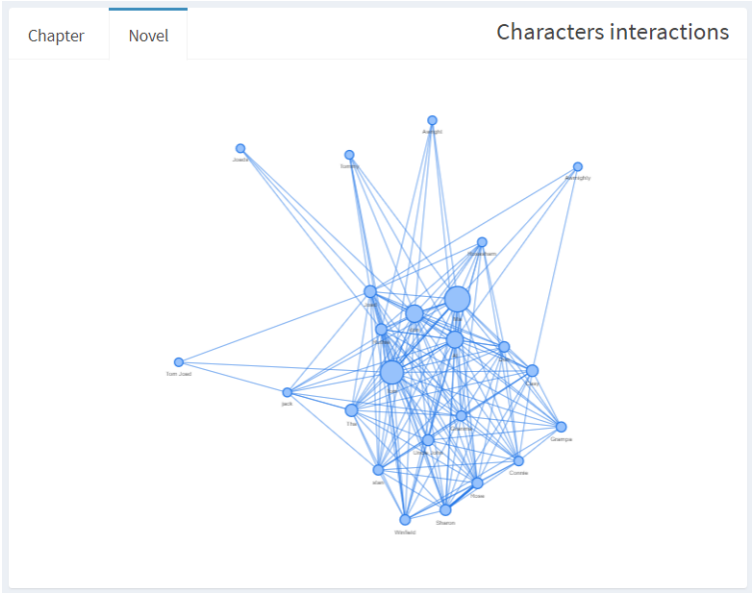


Figura 4.5: Rede de relacionamento entre os personagens ao longo da obra

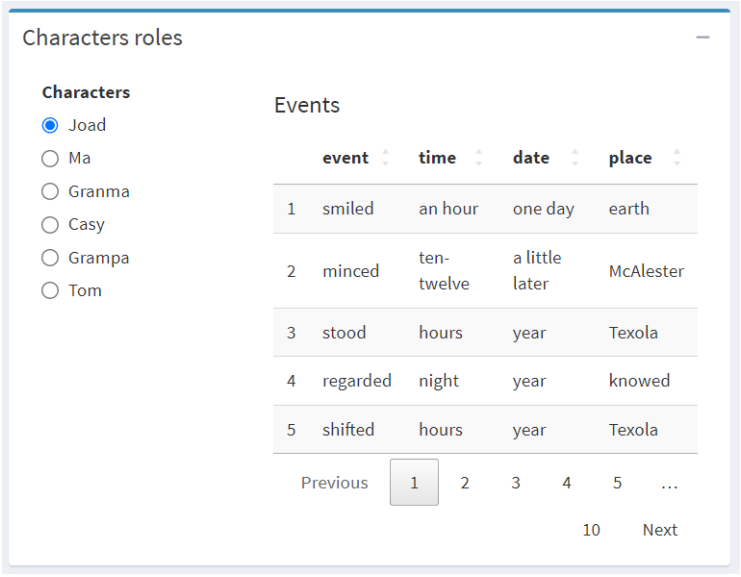


Figura 4.6: Lista das ações dos personagens

4.2 Datas, Marcos temporais e Locais

Nesta secção iremos explorar como a *dashboard* auxilia o leitor a compreender o contexto temporal, de lugar e data na análise de obras literárias. Esses elementos desempenha um papel vital na caracterização dos personagens, na criação de atmosferas distintas e na contextualização dos temas abordados.

O utilizador pode escolher analisar as entidades temporais, de localização e datas, utilizando dois tipos distintos de gráficos. A figura 4.7 e a figura 4.8 ilustram as opções disponíveis ao

utilizador. Pelo menu de navegação, que se encontra na parte superior do quadro, é possível escolher entre as entidades temporais, de data e de localização. Abaixo deste menu, o utilizador tem disponível outro menu de navegação, podendo escolher entre duas visualizações diferentes. Pela opção *X-ray* do menu, o utilizador tem acesso ao gráfico representado na figura 4.7 . No eixo de y, estão representadas as entidades identificadas com o tipo escolhido pelo utilizador. Já no eixo do x, estão representados os números dos capítulos. Através deste gráfico, o utilizador consegue visualizar as entidades identificadas ao longo da obra e os capítulos em que eles são referidos. Já o gráfico da figura 4.8, pode ser visualizado escolhendo a opção *Details*. Neste gráfico estão presentes todas as entidades identificadas na obra. O tamanho de cada um dos círculos é proporcional ao número de capítulos em que a entidade é mencionada.

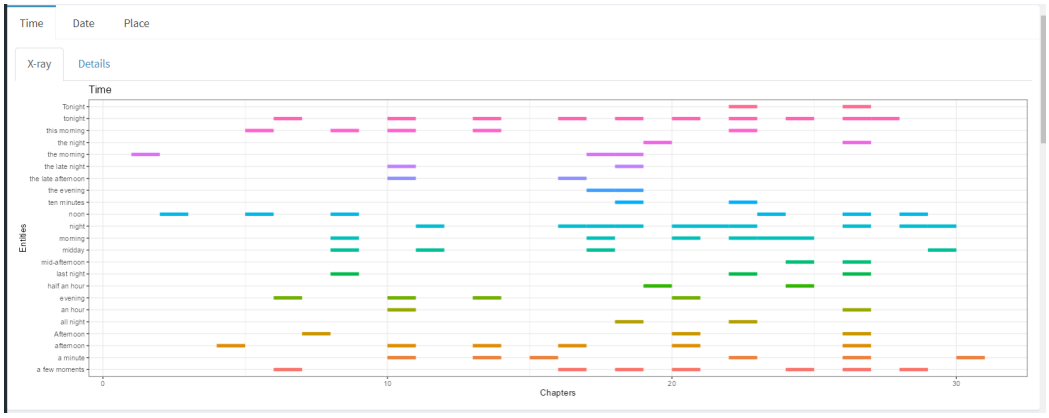


Figura 4.7: Gráfico x-ray dos identificadores temporais identificados ao longo da obra literária

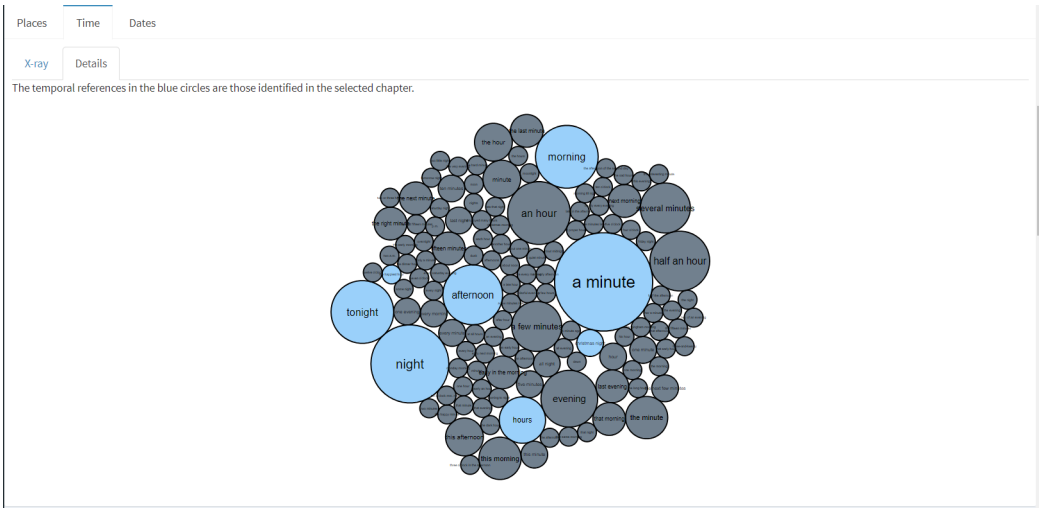


Figura 4.8: Gráfico de detalhes dos identificadores temporais identificados ao longo da obra literária

4.3 Analise de sentimentos e emoções

A analise de sentimentos e emoções são duas ferramentas importantes para se entender a atmosfera e os rumos de uma história literária. Na *dashboard*, o utilizador tem acesso a uma versão do gráfico de linhas apresentado na figura 4.9, a depender da obra que ele tenha escolhido explorar na *dashboard*. No gráfico é representado a evolução do sentimento dos personagens ao longo da obra. No eixo do y temos representados um *score* numérico atribuído aos sentimentos, e no eixo do x é contabilizado o número de frases processadas. A área do gráfico sombreada à azul, marca a região do gráfico referente ao capítulo selecionado para analise.

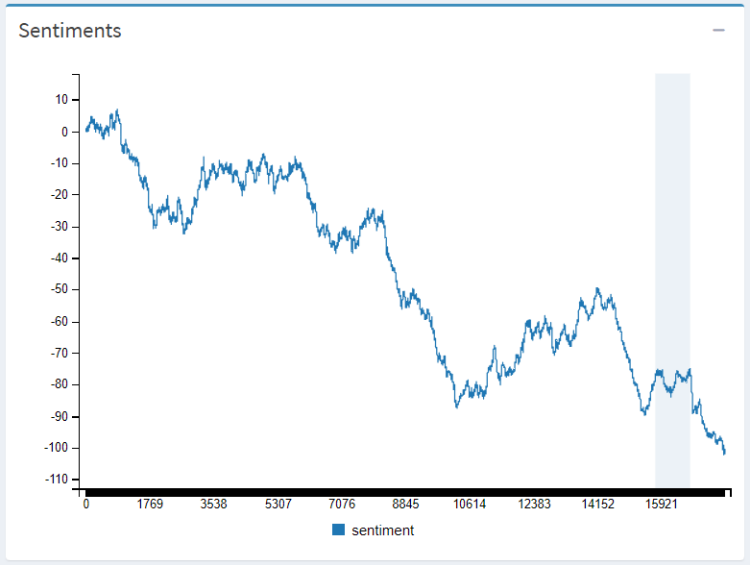


Figura 4.9: Evolução do sentimento ao longo da obra literária

Assim como em outras visualizações da *dashboard*, em que o utilizador pode escolher entre visualizar apenas as informações referentes ao capitulo por ele selecionado ou a todos os capítulos da obra. Ao analisar o gráfico referente as emoções, o utilizador pode escolher entre essas duas visões, através das *tabs* ilustrado na figura 4.10. Caso ele escolha a opção *"Chapter"*, é lhe apresentado o gráfico de linhas da figura 4.10. O gráfico, é utilizado para apresentar os *scores* de 10 emoções diferentes. Cada uma das 10 emoções é representada por três barras distintas no gráfico. Essas três barras correspondem ao capítulo selecionado pelo utilizador, aos capítulos anterior e posterior a este. Esta abordagem permite que os utilizadores tenham uma visão abrangente das variações emocionais ao longo da narrativa. Ao comparar as pontuações emocionais dos capítulos anteriores, atuais e posteriores, podemos destacar mudanças significativas no estado emocional dos personagens.

O utilizador pode também escolher a opção *Novel* e visualizar a variação dos *scores* das emoções ao logo de todos os capítulos, representados no eixo do x. Na figura 4.11 mostramos como o utilizador pode escolher entre visualizar todas as 10 emoções, representado no quadro esquerdo da figura, ou escolher apenas um subconjunto de emoções, ilustrado no quadro direito. Com isso é possível fazer tanto uma avaliação mais detalhada, como estudar as tendências gerais.

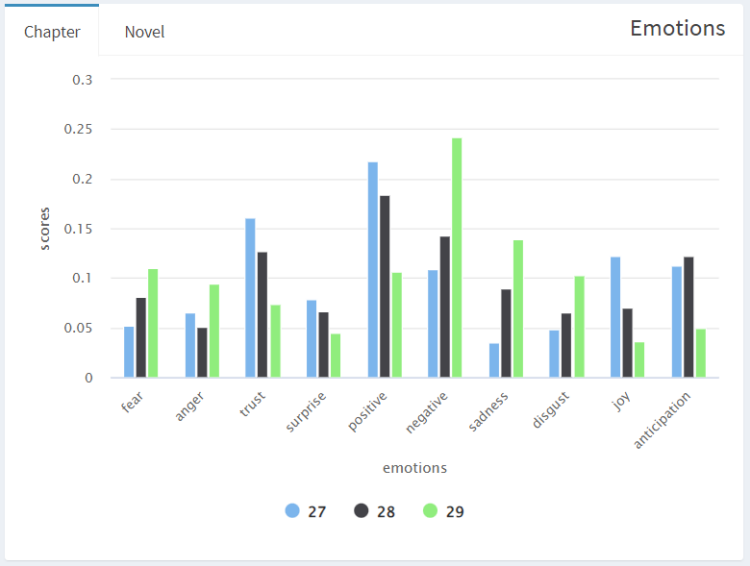


Figura 4.10: As emoções identificadas no capitulo em analise e nos capítulos anterior e posterior a este

Figura 4.11: Evolução das emoções ao longo da obra literária

4.4 Modelação de tópicos

Os tópicos identificados em cada um dos capítulos das diferentes obras, são apresentadas em uma tabela. A figura 4.12 mostra um exemplo de uma dessas tabelas presente na *dashboard*. Para cada um dos tópicos, são identificadas seis palavras que o descrevem.

4.5 Deteção de Keywords

Para cada um dos capítulos são identificados dez *keywords*. O utilizador pode escolher entre *unigrams*, ilustrados no quadro esquerdo da figura 4.13, ou *bigrams*, ilustrados no quadro direito da figura 4.13. Cada *keywords* é associada a um *score* de acordo com a sua relevância.

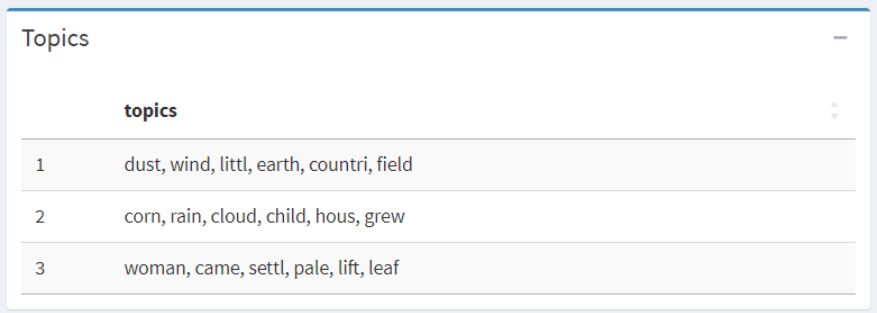


Figura 4.12: Lista dos tópicos identificados no capítulo em análise]

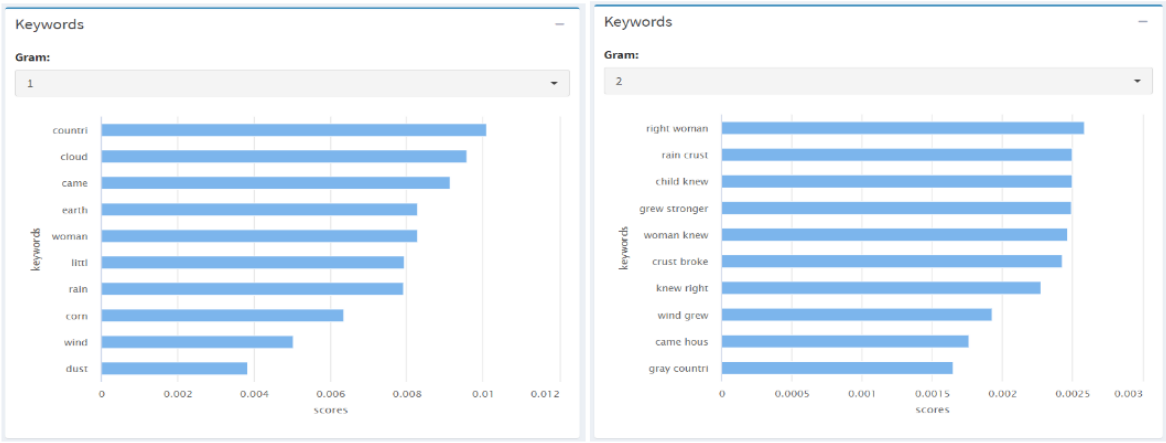


Figura 4.13: As *keywords* e os *scores* associados identificadas no capítulo em análise

As diferentes componentes da *dashboard* foram desenvolvidas com o objetivo de oferecer uma visão abrangente dos diversos elementos que compõem uma obra literária, funcionando assim como uma ferramenta auxiliar no processo de compreensão e caracterização de uma obra literária.

Capítulo 5

Estudo de casos

Neste capítulo, examinaremos como o sistema desenvolvido se comportou ao analisar cinco obras literárias distintas avaliando sua capacidade de adaptação a diferentes estilos, géneros e estruturas narrativas.

5.1 As obras literárias

Nesta secção, faremos uma contextualização e breve descrição das obras literárias que foram escolhidas para análise neste projeto. Estas obras representam um conjunto diversificado de títulos da literatura clássica, que apesar de terem sido lançados em países diferentes, compartilham um período histórico comum, o século XX, com uma variação de apenas algumas décadas entre os lançamentos. As obras escolhidas, possuem diferentes géneros literários, o que permite investigar como diferentes estilos narrativos respondem as técnicas empregadas neste trabalho.

5.1.1 The grapes of wrath

"The Grapes of Wrath"(em português, "As Vinhas da Ira") é um romance clássico escrito pelo autor norte-americano John Steinbeck, em 1939. O livro venceu o prémio Pulitzer de Ficção em 1940 e tem uma grande relevância histórica e social. Esta obra é constituída por 30 capítulos e 213.287 palavras. A história decorre durante a Grande Depressão nos Estados Unidos e segue a família Joad no seu percurso de mudança do estado de Oklahoma, para a Califórnia. Esta é uma mudança forçada pelas duras condições económicas e mudanças no sistema de produção agrícola. Ao longo da obra acompanhamos os desafios e adversidades que os migrantes sofrem ao longo do percurso, assim como a exploração e preconceito de que são alvo ao chegar â Califórnia. O autor utiliza a história da família, para explorar questões sociais, económicas e trabalhistas da época. A narrativa alterna entre a história principal da família Joad e capítulos explanatórios onde são exploradas questões sociais, económicas e históricas ligadas à Grande Depressão e à migração em massa para a Califórnia. Os próprios personagens oferecem uma visão mais ampla do contexto da

época e reforçam os temas centrais do livro. O uso de simbolismo é uma característica marcante desta obra.

5.1.2 Little women

"Little Women"(em português, "Mulherzinhas") é um romance da autora norte-americana Louisa May Alcott, publicado em 1868. O livro possuí 47 capítulos e 196.961 palavras. O romance conta a historia da família March, focando-se na vida de quatro irmãs, Meg, Jo, Beth e Amy, que vivem em Massachusetts, nos Estados Unidos. Assim como "The grapes of wrath", este obra também se passa num importante período histórico, durante a Guerra Civil Americana. Nesta obra são apresentados através das personagens principais, os desafios e sonhos das jovens mulheres da época. Dentre os temas abordados, podemos destacar, o amor fraternal e a família, a autonomia feminina, a moralidade e a educação, a superação de desafios e amadurecimento e o romance.

5.1.3 Lord of the flies

"Lord of the Flies"(em português, "O Senhor das Moscas") é um romance escrito pelo autor britânico William Golding e publicado em 1954. Desenvolve-se ao longo de 12 capítulos, sendo consideravelmente mais curto que os outros, tendo somente 66.187 palavras. A obra ainda é objeto de estudo em muitas escolas e universidades, sendo aconselhado em Portugal no Plano Nacional de Leitura. A história segue um grupo de meninos britânicos, que sobrevivem a um acidente de avião durante uma guerra que não é especificada e acabam presos numa ilha deserta. Neste ambiente hostil, os meninos enfrentam muitos desafios enquanto tentam construir uma sociedade organizada. A obra explora aspetos sombrios da natureza humana, sugerindo que, na ausência de regras e de autoridade externa, as pessoas são capazes de cometer atos cruéis e violentos. Este livro é considerados por muitos uma alegoria política, que critica o autoritarismo e as tendências à violência das sociedades humanas.

5.1.4 1984

"1984"é uma obra do autor britânico George Orwell, publicada em 1949. A obra é uma das mais conhecidas e influentes do género distopia. Ao longo dos 23 capítulos, e 103.950 palavras, o autor apresenta uma sociedade sombria e levanta discussões sobre política e sociedade. O protagonista da história, Winston Smith, vive numa sociedade fictícia, num regime totalitário, no Estado conhecido por Oceania. Nesta realidade, o governo exerce um controlo absoluto sobre a vida dos cidadãos. Winston é um funcionário que trabalha no "Ministério da Verdade", que o governo utiliza para reescrever a história de forma a se adequar às suas narrativas. "1984"faz diversas criticas sociais, entre elas ao totalitarismo e o controle absoluto do Estado sobre a vida das pessoas, destacando temas como a vigilância constante, a manipulação da verdade, a solidão e perda dos direitos humanos.

5.1.5 The Godfather

"The Godfather"é um romance policial escrito por Mario Puzo. Publicado em 1969, a obra tornou-se uma das mais influentes do século XX, tendo dado origem a uma série de filmes de sucesso. A história é contada em 32 capítulos e 184.247 palavras. O livro também é conhecido por sua riqueza na caracterização e desenvolvimento dos seus personagens. São explorados os altos e baixos da vida da máfia nos Estados Unidos, através da família Corleone, uma poderosa família italo-americana, envolvida no crime organizado. O personagem principal da história é o patriarca da família e chefe do crime, o respeitado e temido Don Vito Corleone. O autor apresenta complexas relações familiares, histórias de traição, vingança e rivalidade entre gangues. Através destas narrativas, são abordados temas como honra, poder, lealdade e moralidade no crime organizado. O livro ainda levanta importantes questões sobre imigração e identidade cultural.

5.1.6 A sumarização dos capítulos

O objetivo de incluir a sumarização das obras foi o de fornecer ao leitor um breve resumo dos acontecimentos mais relevantes de cada um dos capítulos e a contextualização da obra, além de explicar o contexto de onde se passa a história. Porém, após uma extensiva análise dos resumos de cada uma das obras, compreendemos que estes resumos não oferecem uma clara visão dos acontecimentos ao longo das histórias. Não obstante que, em algumas das obras, eles ajudam o leitor a compreender outros aspetos da historia, como o contexto social, o meio onde se passa a história e até mesmo características de alguns personagens. Vale ressaltar que os resumos apresentados na *dashboard*, utilizam o método extrativo, nesta abordagem o resumo é composto por sentenças retiradas diretamente do texto original. As sentenças selecionadas são as consideradas as mais importantes com base em critérios como a pontuação de importância calculada pelo algoritmo TextRank.

Na obra "The grapes of wrath"temos cenários distintos. Esta obra intercala capítulos onde são contadas as história da família principal com capítulos explanatórios. Nos capítulos explanatórios, os resumos são mais eficientes, conseguindo transmitir partes do tema retratados pelo autor. Como por exemplo no texto da figura 5.1, é apresentado autoestrada 66, descobrimos que ela é utilizada por refugiados que procuram melhores condições de vida na Califórnia. Sendo que a migração é um dos temas centrais da história. Aqui é importante ressaltar que o autor faz uso de muitos simbolismo e estes não são facilmente capturados pelos algoritmos de *NLP* usados.

Por outro lado, em grande parte dos capítulos em que é contada a história da família, os resumos são formados por diálogos entre personagens ou frases que sem contexto adicional não ajudam o leitor a compreender a história, como acontece no texto ilustrado na figura 5.2.

O problema acima citado é transversal a todas as obras. Por exemplo na obra "The Godfather", encontramos outros resumos que assim como o da figura 5.3, falam sobre uma ação ou pensamento, mas não identificam o personagem que as praticam ou a quem se direcionam.

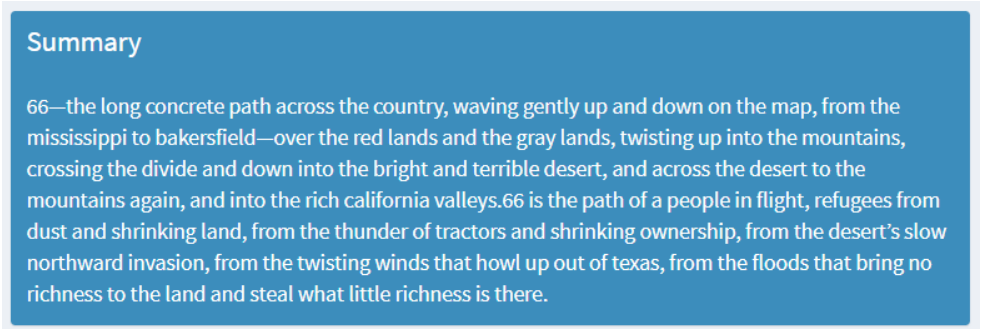


Figura 5.1: Resumo de um capítulo da explanatório de The grapes of wrath

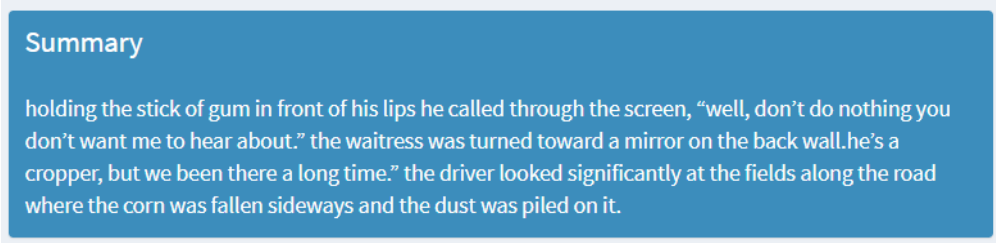


Figura 5.2: Resumo de um capítulo da história de The grapes of wrath]

Nesta obra também encontramos resumos mais interessantes, como o da figura 5.4, que oferecem informações importantes ao leitor sobre os acontecimentos e o ambiente da história, nesta passagem conseguimos ver indícios de ligação com a máfia, através da menção do personagem deixar de ser o "Caporegime", um membro de alto escalão em uma família ou organização criminosa. Em outras passagens como "The way to do it would be to have him heavily implicated so that it's not an honest police captain doing his duty but a crooked police official mixed up in the rackets who got what was coming to him, like any crook.", conseguimos identificar temas como corrupção. Os resumos desta obra, também fazem alusão a temas como vingança e crime organizado.

Alguns dos resumos de "Little women"permitem conhecer um pouco mais das características das personagens; isso é ilustrado no resumo da figura 5.5, ou até mesmo acontecimentos

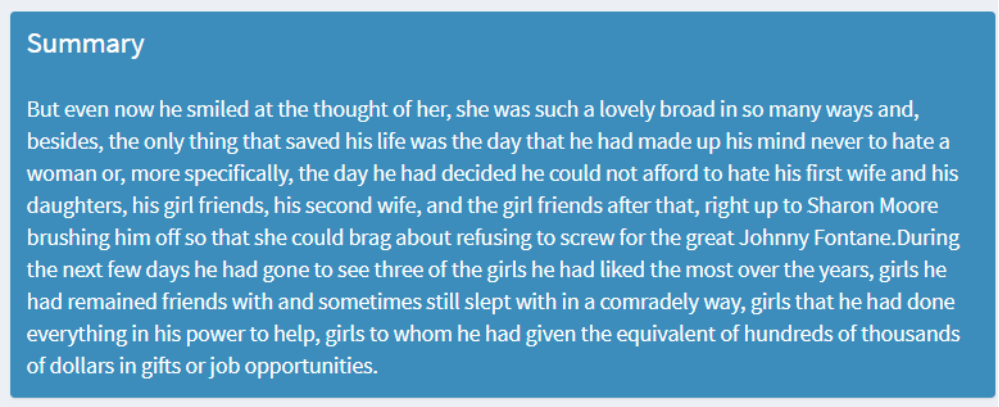


Figura 5.3: Resumo de um capítulo da obra The Godfather

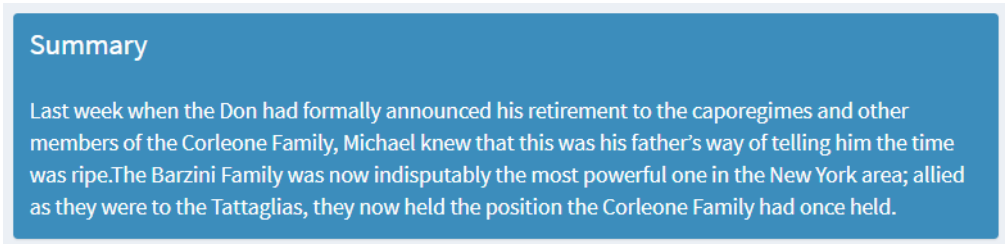


Figura 5.4: Resumo de um capítulo da obra The godfather

importantes da história, como o amadurecimento da personagem Meg mostrado no final do texto da figura 5.6.

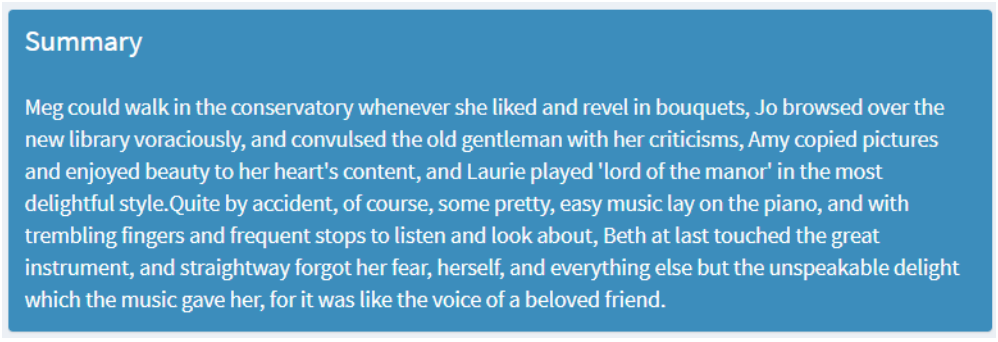


Figura 5.5: Resumo de um capítulo da obra Little women, mostrando algumas características das personagens

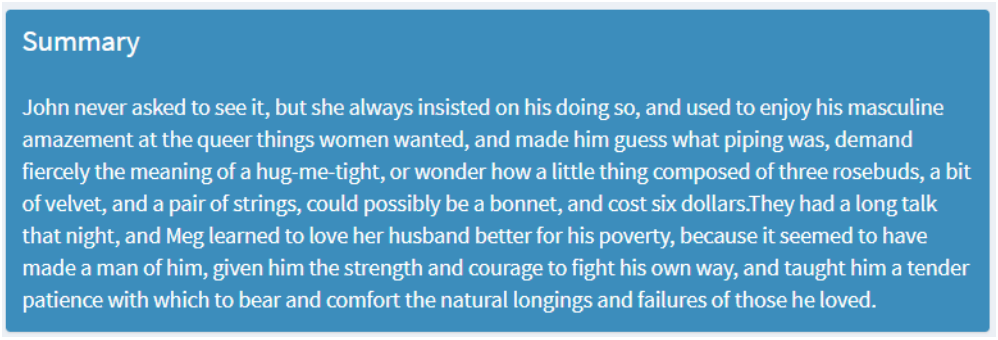


Figura 5.6: Resumo de um capítulo da obra Little women que destaca um acontecimento importante para a história

Os resumos da obra "1984", foram os mais bem-sucedidos dentre as obras em análise. Os textos permitem ter uma visão de alto nível do cenário social e político em que a história se desenvolve, e também fornecem pistas do descontentamento do personagem principal com a situação política e social. Os resumos apresentam também alguns acontecimentos marcantes, como de tortura, do texto da figura 5.7, que serve também para mostrar ao leitor as atrocidades extremas que partido político da obra comete.

O facto de "1984", possuir uma estrutura narrativa mais descritiva, onde o autor descreve em muitos detalhes os personagens e o ambiente em que eles estão inseridos, podem ter ajudado

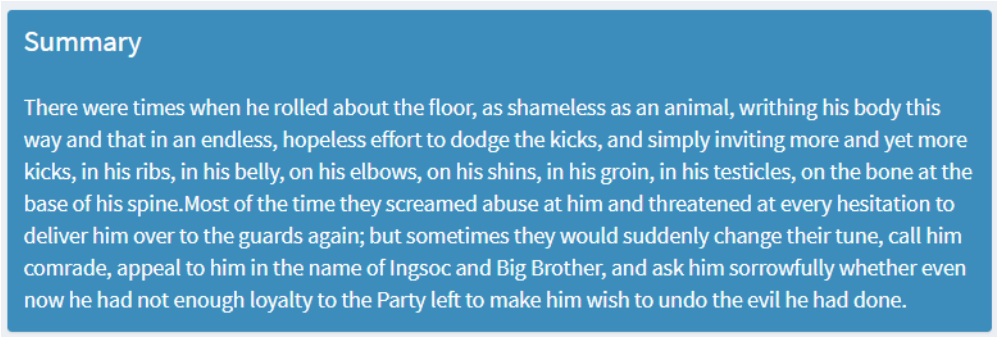


Figura 5.7: Resumo que destaca um momento importante para o personagem principal da obra 1984

o algoritmo de sumarização a obter melhores resultados. "Little women"também poderá ter se beneficiado de ter uma estrutura narrativa mais simples. Por outro lado "The grapes of wrath", possui uma estrutura mais complexa, carregada de simbolismos e capítulos mais longos, fatores que dificultam a tarefa de sumarização automática utilizando o método extrativo.

5.2 Personagens

Verificamos que o sistema desenvolvido identifica com sucesso todos os personagens principais nas obras em análise. No entanto, apresenta algumas limitações, como ilustrado na figura 5.8, que representa os personagens identificados na obra "The Godfather". Os quadrados amarelos destacam alguns personagens que merecem análise mais detalhada.

Primeiramente, observamos que o Named-entity recognition (NER) identifica "don"e "don corleone"como entidades separadas, embora se refiram à mesma pessoa. O mesmo padrão ocorre com "michael"e "michael corleone", assim como "sonny"e "sonny corleone". O algoritmo não consegue fazer a conexão que esses nomes se referem à mesma pessoa. Além disso, o algoritmo erroneamente identifica "corleone"como uma pessoa, quando, na verdade, é o sobrenome da família. Esses problemas de identificação repetem-se em outras obras, como "The Grapes of Wrath", onde o NER também confunde nomes de personagens com seus sobrenomes, e em "Little Women", onde o sobrenome da família é identificado como uma pessoa.

O modelo utilizado para identificar essas entidades é pré-treinado e, embora produza resultados satisfatórios em muitos casos, a identificação de personagens em uma obra literária é uma tarefa desafiadora. Isso ocorre porque é um problema altamente dependente de contexto. Em diferentes partes do texto, um personagem pode ser referido de maneiras distintas, como "don"ou "don corleone". Deste modo, o modelo pode não ter informações contextuais suficientes para entender que essas variações se referem à mesma pessoa.

Na figura 5.9, conseguimos facilmente identificar que alguns personagens estão conectados com uma maior número de personagens e possuem ligações mais fortes. Os gráficos de redes, auxiliam não só na identificação das relações entre personagens, como também na identificação

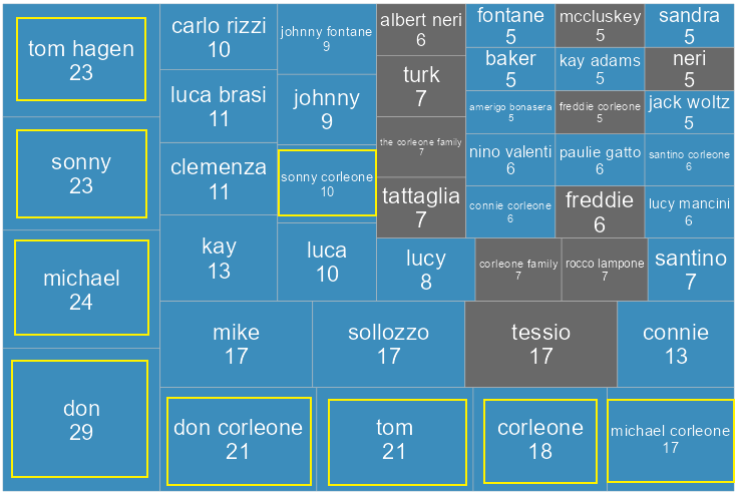


Figura 5.8: Personagens identificados na obra The godfather

dos personagens centrais da história, através do tamanho dos círculos associados aos personagens, e do número e da espessura das suas conexões. Esta visualização foi bem sucedida em todas as obras analisadas.

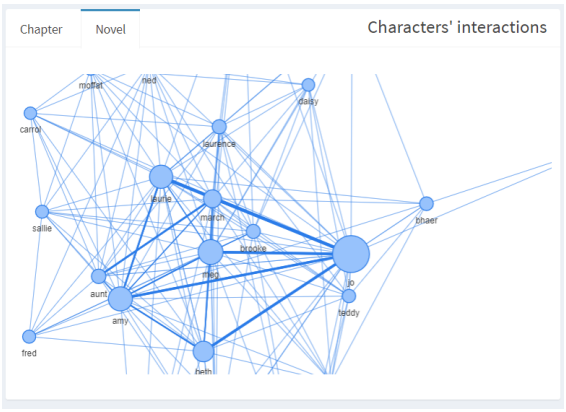


Figura 5.9: Gráfico de eventos identificados em um capítulo

5.3 Datas, Marcos temporais e Locais

Um dos aspetos importantes a identificar nas obras, são as referencias temporais, as de data, e as de local. De entre essas, a que obteve melhor resultado, foi a identificação de entidades do tipo local. Através destas entidades conseguimos identificar as principais localizações e definir a sua relevância na obra através do gráfico de entidades de localização (exemplo, o da figura 5.10). Através da figura 5.11, é possível identificar alguns dos desafios relacionados às entidades temporais e de data. Tomando como exemplo o evento "disappear"associado ao personagem "Sonny", podemos observar que temos informações limitadas, sabendo apenas que ele desapareceu, possivelmente durante a Páscoa ("Easter"). No entanto, as referências "half hour"e "years"não fornecem detalhes suficientes para fazer inferências significativas sobre o evento. Isso

se repete noutras visualizações. As ações dos personagens, as referências temporais, de data, e de local muitas vezes são expressas de maneira indireta e simbólica, tornando difícil para algoritmos identificá-las com precisão.

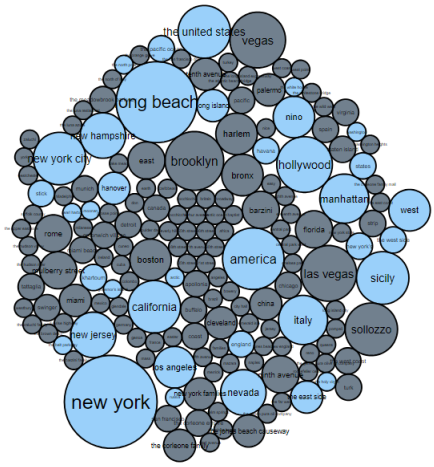


Figura 5.10: Gráfico de localizações identificadas na obra The godfather

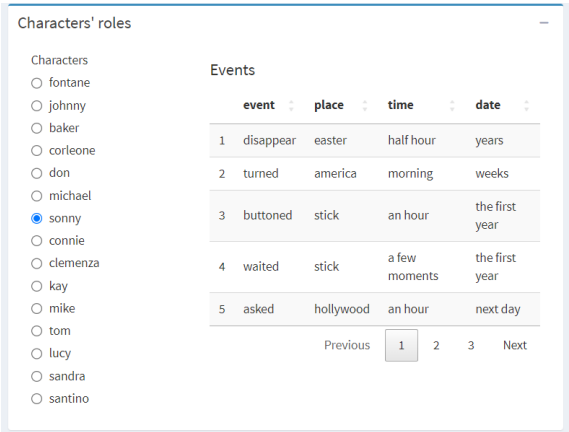


Figura 5.11: Gráfico de ações dos personagens

5.4 Análise de sentimentos e emoções

Os gráficos de sentimentos e emoções, conseguiram captar o tom emocional das obras analisadas, identificando pontos importantes da história, que se destacam por variações significativas nos gráficos. No gráfico da figura 5.12, conseguimos identificar uma queda acentuada no capítulo sinalizado (representado com fundo a cinzento). Essa queda se reflete no gráfico de emoções da Figura 5.13, onde o valor do sentimento negativo supera o positivo. Podemos também notar que a emoção predominante é o medo, seguido pela tristeza. Este é um dos exemplos em que estes gráficos conseguem capturar acontecimentos importantes, neste capítulo o personagem principal da obra "1984"é torturado.

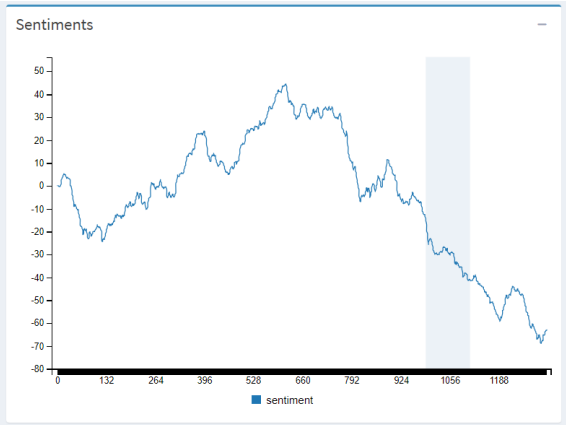


Figura 5.12: Análise de sentimentos de um capítulo da obra 1984

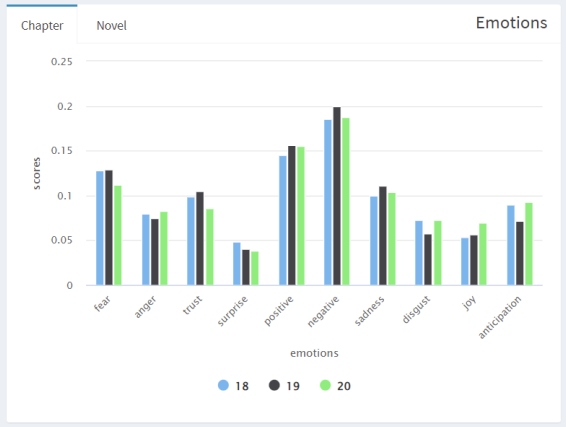


Figura 5.13: Análise das emoções de um capítulo da obra 1984

5.5 Topic modeling e extração de keywords

As técnicas de *topic modeling* e extração de *keywords*, não obtiveram sucesso na identificação dos principais temas e tópicos na maioria dos capítulos analisados. Muitas das *keywords* identificadas consistiram de nomes de personagens e de lugares, que eventualmente não ajudam o leitor na compreensão da obra. Os tópicos enfrentam um problema similar, um desafio semelhante, sendo frequentemente compostos por nomes de personagens ou verbos. Embora tenhamos empregado a métrica *coehrence* para determinar o número ideal de tópicos a cada capítulo, ainda enfrentamos o desafio de lidar com a redundância entre muitos desses tópicos.

Ao utilizarmos a *dashboad* para analisar a obra "the godfather", a *keyword* "famili"aparece entre as *keywords* identificadas em muitos capítulos, fazendo alusão ao tema família, que é um dos temas centrais da trama de Mário Puzo.

"1984"foi a obra em que o algoritmo de extração de *kewords*, obteve o melhor resultado pois conseguiu identificar um subconjunto maior de *keywords* relacionadas a um dos temas principais da obra, a política. Entre as *keywords* identificadas, destacamos: "Party members",

"Party loyalty", "Party", "Poor people", "Party doctrine", "Party propaganda"que, apesar de não fornecerem detalhes do cenário político, ajudam o leitor a compreender que a política é um dos temas centrais. Para além disso, o algoritmo também identifica *keywords* que dão pistas ao leitor sobre o clima da trama, por exemplo temos as *keywords*: "war", "power", "pain", que fazem alusão a um clima de tensão. É interessante notar, ainda analisando o comportamento do sistema em relação a obra 1984, as *keywords*, "pain", "prison", "confess", aparecem nos capítulos em que o personagem principal sofre tortura, alinhado com a descrição da tortura presentes no resumo da figura5.7 e uma queda acentuada no sentimento representado na figura 5.12.

Capítulo 6

Conclusões

Neste trabalho, desenvolvemos um sistema que tem por objetivo ser uma ferramenta auxiliar à análise literária tradicional. Se propondo a oferecer um panorama geral e alto nível de elementos genéricos de uma obra literária, permitindo a identificação de personagens principais, pontos onde temos uma mudança no tom emocional da obra, etc. Tem como publico alvo leitores que tem a necessidade de fazer analises literárias comparativas ou procuram uma rápida visão alto nível e introdutória às obras literárias. Para desenvolver tal sistema, foi necessária a aplicação e adaptação de técnicas de Processamento de linguagem natural (NLP), disponibilizadas por bibliotecas públicas, no cenário aplicado em estudo. As técnicas de NLP em estudo são: a *topic modeling*, Named-entity recognition (NER), a extração de *keywords*, a analise de sentimentos, a deteção de emoções e a sumarização de texto. Por fim, foi criada uma *dashboard* com o objetivo de apresentar ao utilizador uma representação visual das análises sobre os diferentes aspetos das obras literárias.

Podemos apontar duas contribuições-chave deste trabalho, que foram alcançadas durante seu processo de desenvolvimento:

- Exploração e aplicação de técnicas NLP, disponibilizadas por bibliotecas públicas da biblioteca *python*. Para realização desta tarefa, foi necessário, estudar as limitações destas bibliotecas e desenvolver alternativas para melhor tentar explorar os seus pontos positivos e limitar os impactos dos problemas identificados. Temos como produto final desta etapa, um conjunto de *scripts*, que automaticamente aplicam as técnicas de NLP em estudo. O utilizador apenas precisa inserir um arquivo no formato "txt"da obra que deseja estudar, e esse arquivo deve incluir a marcação das divisões dos capítulos;
- O desenvolvimento de uma *dashboard*, que permite ao utilizador explorar diferentes aspetos de um conjunto de obras literais, através de representações gráficas. Através desta ferramenta, o utilizador pode fazer comparações entre obras ou explorar uma única obra capítulo à capítulo.

De entre as tarefas que o sistema se propõe a executar, gostaríamos de destacar três em que

ele obteve resultados muito satisfatórios:

- O sistema conseguiu identificar os personagens principais em todas as obras analisadas no estudo de caso em questão. Para além disso, através dos gráficos o utilizados consegue facilmente identificar quais os personagens com maior relevância para a história. O sistema consegue eficientemente também apresentar ao utilizador as principais ligações entre personagens e a intensidade dessas ligações;
- As referências de localização, cumprem o seu papel e conseguem identificar os principais lugares em que se desenrola a historia;
- As técnicas de análise de sentimentos e deteção de emoções, possibilitaram a criação de gráficos indicativos do tom emocional das narrativas e permitem identificar pontos importantes de variação positiva e negativa do sentimento ao longo da história.

As obras literárias apresentam estruturas complexas, os autores fazem uso de diferentes recursos linguísticos para contar uma história, sendo que alguns deles como o simbolismo, a ambiguidade, metáforas, entre outros, podem não ser facilmente identificados por meio das técnicas de NLP atualmente disponíveis. Além disso, as obras literárias frequentemente dependem de contextos intrínsecos à narrativa, o que torna desafiante a interpretação por métodos automáticos.

Essas características mencionadas contribuem para que o sistema não desempenhe tão bem em algumas das tarefas que lhe foram atribuídas:

- A modelação de tópicos e a extração de *keywords* não conseguiram identificar de maneira precisa os principais temas e tópicos abordados na narrativa. A frequente menção de nomes de personagens, por exemplo, pode ter causado confusão nos algoritmos, levando-os a considerar esses nomes como tópicos ou palavras-chave;
- A identificação de personagens, apesar de no geral apresentar um resultado muito bom, tem alguns problemas. Por exemplo, em uma das obras o sistema identifica "Don"e "Don Corleone"como sendo duas pessoas distintas, quando na verdade são a mesma pessoa, também identifica erroneamente o sobrenome "Corleone"como uma pessoa. Este problema repete-se em mais de uma obra literária. Problemas de identificação de contexto, aliados ao facto que um mesmo personagem pode ser mencionando diferentes formas em partes diferentes da narrativa, podem estar na raiz desse problema;
- O sistema acaba também por ter lacunas na identificação de entidades do tipo temporal e de data. Numa narrativa, muitas das vezes referências temporais e de data são feitas de formas subjetivas utilizando nuances das linguísticas, ou utilizando expressões relativas como "há uma semana", "no mês passado", que sem contexto adicional não ajudam o utilizador do sistema a compreender a passagem de tempo na narrativa e construir uma ordem cronológica dos acontecimentos.

Deixamos também algumas sugestões de trabalhos futuros para aprimorar ainda mais o sistema:

- Explorar e integrar técnicas de NLP mais avançadas que possam lidar com as nuances linguísticas, simbolismo, ambiguidades e a forte dependência de contexto presentes em obras literárias;
- Implementar melhorias nas partes do sistema relacionadas à *topic modeling* e à extração de *keywords*. De forma a conseguirmos distinguir com mais eficiência entre nomes de personagens e temas reais, aprimorando assim a identificação de tópicos relevantes na narrativa;
- Desenvolver algoritmos mais robustos para identificar entidades temporais e de data, levando em consideração o contexto intrínseco à narrativa. Isso ajudará a construir uma representação cronológica mais precisa dos eventos nas obras literárias;
- Desenvolver um método mais robusto para identificação de eventos nas obras;
- Abordar o desafio das múltiplas referências a um mesmo personagem ao longo da narrativa. Isso pode envolver a implementação de técnicas avançadas de resolução de correferência para garantir uma identificação unificada dos personagens;
- Expor o sistema a diferentes grupos de utilizador e recolher o *feedback* desses grupos, para através deles identificar pontos de melhoria de usabilidade e questões técnicas.
- Fazer uma análise de usabilidade do sistema, em particular da *dashboard*.

Bibliografia

[1] Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7): e12189, 2020.

[2] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth Trippe, Juan Gutiérrez, and Krys Kochut. Text summarization techniques: A brief survey. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 8:397–405, 07 2017. doi:10.14569/IJACSA.2017.081052.

[3] Vimala Balakrishnan and Lloyd-Yemoh Ethel. Stemming and lemmatization: A comparison of retrieval performances. *Lecture Notes on Software Engineering*, 2:262–267, 01 2014. doi:10.7763/LNSE.2014.V2.134.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[5] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. Yake! collection-independent automatic keyword extractor. In *Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings 40*, pages 806–810. Springer, 2018.

[6] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289, 2020. ISSN: 0020-0255. doi:https://doi.org/10.1016/j.ins.2019.09.013.

[7] K.R. Chowdhary. *Fundamentals of Artificial Intelligence*. 02 2020. ISBN: 978-81-322-3970-3. doi:10.1007/978-81-322-3972-7.

[8] Cach Dang, María Moreno García, and Fernando De La Prieta. Sentiment analysis based on deep learning: A comparative study. *Electronics*, 9:483, 03 2020. doi:10.3390/electronics9030483.

[9] Ran Ding, Ramesh Nallapati, and Bing Xiang. Coherence-aware neural topic modeling. *arXiv preprint arXiv:1809.02687*, pages 830–836, 2018. doi:10.18653/v1/D18-1096.

[10] Wafaa S. El-Kassas, Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679, 2021. ISSN: 0957-4174. doi:https://doi.org/10.1016/j.eswa.2020.113679.

[11] Mahak Gambhir and Vishal Gupta. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47:1–66, 2017. doi:10.1007/s10462-016-9475-9.

[12] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78:15169–15211, 2019.

[13] Yue Kang, Zhao Cai, Chee-Wee Tan, Qian Huang, and Hefu Liu. Natural language processing (nlp) in management research: A literature review. *Journal of Management Analytics*, 7: 1–34, 05 2020. doi:10.1080/23270012.2020.1756939.

[14] Subbu Kannan, Vairaprakash Gurusamy, S Vijayarani, J Ilamathi, Ms Nithya, S Kannan, and V Gurusamy. Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks*, 5(1):7–16, 2014.

[15] Anne Kao and Stephen Poteet. *Natural Language Processing and Text Mining*, volume 7. 01 2007. doi:10.1145/1089815.1089816.

[16] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural language processing: State of the art, current trends and challenges. *Multimedia tools and applications*, 82(3):3713–3744, 2023.

[17] Divya Khyani, BS Siddhartha, NM Niveditha, and BM Divya. An interpretation of lemmatization and stemming in natural language processing. *Journal of University of Shanghai for Science and Technology*, 22(10):350–357, 2021.

[18] Jing li, Aixin Sun, Ray Han, and Chenliang Li. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, PP:1–1, 03 2020. doi:10.1109/TKDE.2020.2981314.

[19] Daniel Maier, Annie Waldherr, Peter Miltner, Gregor Wiedemann, Andreas Niekler, Alexa Keinert, Barbara Pfetsch, Gerhard Heyer, Ueli Reber, Thomas Häussler, et al. Applying lda topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, 12(2-3):93–118, 2018.

[20] Mónica Marrero, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato, and Juan Miguel Gómez-Berbís. Named entity recognition: Fallacies, challenges and opportunities. *Computer Standards Interfaces*, 35(5):482–489, 2013. ISSN: 0920-5489. doi:https://doi.org/10.1016/j.csi.2012.09.004.

[21] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 2014. ISSN: 2090-4479. doi:https://doi.org/10.1016/j.asej.2014.04.011.

[22] Behrang Mohit and Zitouni Imed. Natural language processing of semitic languages, 2014.

[23] David Newman, Edwin V Bonilla, and Wray Buntine. Improving topic coherence with regularized topic models. 24, 2011.

[24] Aytuğ Onan, Serdar Korukoğlu, and Hasan Bulut. Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications*, 57:232–247, 2016. ISSN: 0957-4174. doi:https://doi.org/10.1016/j.eswa.2016.03.045.

[25] Daniel W Otter, Julian R Medina, and Jugal K Kalita. A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, 32(2):604–624, 2020.

[26] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. *Automatic Keyword Extraction from Individual Documents*, pages 1 – 20. 03 2010. ISBN: 9780470689646. doi:10.1002/9780470689646.ch1.

[27] Sifatullah Siddiqi and Aditi Sharan. Keyword and keyphrase extraction techniques: A literature review. *International Journal of Computer Applications*, 109:18–23, 01 2015. doi:10.5120/19161-0607.

[28] Ashok Srivastava and Mehran Sahami. *Text Mining: Classification, Clustering, and Applications*. Chapman & Hall/CRC, 1st edition, 2009. ISBN: 1420059408.

[29] Ike Vayansky and Sathish Kumar. A review of topic modeling methods. *Information Systems*, 94:101582, 06 2020. doi:10.1016/j.is.2020.101582.

[30] S Vijayarani, Ms J Ilamathi, Ms Nithya, et al. Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, 5(1): 7–16, 2015.

[31] Lei Zhang, Shuai Wang, and Bing Liu. Deep learning for sentiment analysis : A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8, 01 2018. doi:10.1002/widm.1253.

[32] Weizhong Zhao, James J Chen, Roger Perkins, Zhichao Liu, Weigong Ge, Yijun Ding, and Wen Zou. A heuristic approach to determine an appropriate number of topics in topic modeling. In *BMC bioinformatics*, volume 16, pages 1–10. Springer, 2015.