INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Classificação Automática de Registos Eletrónicos Médicos por Diagnóstico

Ana Rita Amaro Barros

Mestrado em Sistemas Integrados de Apoio à Decisão

Orientador:

Doutor João Carlos Amaro Ferreira, Professor Auxiliar

Iscte – Instituto Universitário de Lisboa

Outubro, 2020

iscte

TECNOLOGIAS
E ARQUITETURA

Classificação Automática de Registos Eletrónicos Médicos por Diagnóstico

Ana Rita Amaro Barros

Mestrado em Sistemas Integrados de Apoio à Decisão

Orientador:

Doutor João Carlos Amaro Ferreira, Professor Auxiliar

Iscte – Instituto Universitário de Lisboa

Outubro, 2020

Agradecimentos

Quero agradecer ao meu orientador Dr. João Ferreira pela disponibilidade, apoio aconselhamento prestados ao longo deste trabalho, pois sem esta orientação não teria sido possível este trabalho.

Gostaria também de agradecer a dois dos meus amigos mais próximos Paula Moreira e João Rodrigues pelo apoio, pela motivação e pela paciência que tiveram ao longo destes dois anos.

Por fim, mas não menos importante, gostaria de agradecer à minha família por todo o amor e suporte prestado, que foram sem dúvida cruciais para a conclusão desta etapa tão importante.

Resumo

A crescente implementação de sistemas de registos eletrônicos médicos (REM's) nos Hospitais, com vista a apoiar o atendimento individual dos pacientes, está a provocar um aumento do processamento e armazenamento dos dados clínicos diariamente. Estes registos contêm uma fonte infindável de informação clínica, no entanto o facto de não haver estrutura no texto produzido pelos médicos e o facto das informações introduzidas divergirem de paciente para paciente e de especialidade médica para especialidade médica, dificulta o aproveitamento destes dados. Outra dificuldade que existe na análise deste tipo de dados é conseguir criar um sistema capaz de extrair informação minuciosa presente nos REM's, de forma a ajudar os profissionais de saúde a reduzir a taxa de erro de diagnóstico, prevendo o tipo de doença do paciente. Atualmente, para superar este desafio os hospitais realizam este processo manualmente, no entanto este processo é longo e está suscetível a erros. Esta dissertação pretende propor uma solução para este problema, ao utilizar técnicas de Processamento de Linguagem Natural e de Aprendizagem Automática, de forma a permitir um sistema que possibilite a extração de conhecimento clínico e respetiva classificação do REM por tipo de doença/ diagnóstico, de uma forma automática. Este sistema foi desenvolvido em língua portuguesa, visto que todos os sistemas médicos de extração de conhecimento existentes são desenvolvidos para língua inglesa. Este cenário visa ajudar na evolução do aproveitamento das informações contidas nos REM's e, consequentemente, visa contribuir para o crescimento deste tipo de sistemas dentro do hospital português envolvido nesta dissertação.

Palavras-Chave: Extração de Informação, Extração de Conhecimento, Mineração de Texto, Processamento de Linguagem Natural, Classificação Automática, Classificação Multiclasse

Abstract

The growing implementation of electronic medical record (EMR's) systems in Hospitals, to support individual patient care, is causing an increase in the processing and storage of clinical data daily. These records contain an endless source of clinical information, however, the fact that there is no structure in the text produced by doctors and the fact that the information entered differ from patient to patient and from medical speciality to medical speciality, makes it difficult to use these data. Another difficulty that exists in the analysis of this type of data is to be able to create a system capable of extracting detailed information present in the EMR's, in order to help health professionals to reduce the error rate of diagnosis, predicting the type of disease of the patient. Currently, to overcome this challenge, hospitals carry out this process manually, however, this process is long and susceptible to errors. This dissertation intends to propose a solution to this problem, using techniques of Natural Language Processing and Machine Learning, in order to allow a system that allows the extraction of clinical knowledge and respective classification of EMR by type of disease/diagnosis, from an automatically. This system was developed in Portuguese language since all existing medical knowledge extraction systems are developed for English. This scenario aims to help in the evolution of the use of the information contained in the EMR's and, consequently, aims to contribute to the growth of this type of systems within the Portuguese hospital involved in this dissertation.

Keywords: Information Extraction, Knowledge Extraction, Text Mining, Natural Language Processing, Automatic Classification, Multiclass Classification

Índice Geral

Índice de Tabelas

Índice de Figuras

Glossário de Abreviaturas e Siglas

TM	*Text Mining*
NLP	*Natural Language Processing*
DM	*Data Mining*
AA	Aprendizagem Automática
REM	Registo Eletrónico Médico
BioNLP	*Biomedical Natural Language Processing*
GATE	*General Architecture for Text Engineering*
HITEx	*Helth Information Text Extraction*
cTAKES	*clinical Text Analysis and Knowledge Extration System*
MedLEE	*Medical Language Extraction and Encoding System*
MMTx	*MetaMap Transfer*
SVM	*Support Vector machine*
TF-IDF	*Term Frequency – Inverse Document Frequency*
POS-TAG	*Part-of-speech Tagging*
NLTK	*Natural Language Toolkit*

Capítulo 1 – Introdução

1.1. Enquadramento do Tema

O aumento do uso de técnicas exploratórias que permitem a análise de dados em diversas áreas da sociedade tem vindo a se tornar crucial, visto que possibilita às diferentes organizações adquirir mais conhecimento e consequentemente melhorar os seus processos. A área da saúde, não é uma exceção e tem se observado um crescimento notório no uso destas técnicas no contexto hospitalar [1]. Nos diversos processos diários de um hospital, são recolhidos, armazenados e processados uma grande quantidade de dados quer estejam em formato estruturado ou não [2]. Estes dados são armazenados nas bases de dados dos Hospitais e encontram-se nos Registos Eletrónicos Médicos (REM's), que consistem numa "folha" digital na qual o médico descreve o estado clínico do paciente. Antigamente o médico escrevia numa folha à mão, no entanto com o avanço da tecnologia estes são preenchidos em formato de texto narrativo diretamente no computador, o que possibilita a extração de informação dos mesmos [3]. Apesar de existir uma quantidade enorme de dados presentes nos REM's e de existir a hipótese de processamento dos mesmos, eles estão organizados de forma não estruturada [4] o que dificulta a sua extração. Estes registos podem conter diversas informações, tais como, os sinais vitais dos pacientes, os diagnósticos de exames, os medicamentos e tratamentos médicos.

Atualmente, os Hospitais beneficiam de um sistema de tecnologia de informação clínica, neste caso em específico, um sistema de REM's, pelo que conseguem reduzir o tempo gasto na recolha dos dados clínicos [5]. Além deste benefício, é possível efetuar análises aos dados gerados, como por exemplo, encontrar novos padrões biomédicos [6], criar/ melhorar modelos preditivos de doenças, sugerir recomendações de exames [7], ajudar nas tomadas de decisões clínicas e administrativas e basear os estudos científicos e/ ou a literatura biomédica [6].

Estes sistemas de informação podem ajudar a reduzir os erros médicos, bem como a auxiliar os médicos a determinar um diagnóstico correto. Isto proporciona uma melhor qualidade de atendimento aos pacientes, fornecendo aos profissionais de saúde informação e feedback mais precisa e rápida sobre o estado de saúde dos mesmos [3].

A análise dos dados não estruturados, ou seja, dados em texto narrativo, presentes nos REM's é realizada através de técnicas de *text mining*, pois estas auxiliam a extrair corretamente as informações médicas e apresentá-las de forma estruturada [4].

1.2. *Text Mining*

O *Data Mining* (DM) consiste na procura de padrões em dados estruturados, ao contrário do *Text Mining* (TM) que procura padrões em dados não estruturados, denominados de texto livre. Ambas as técnicas procuram encontrar informações úteis, no entanto as técnicas utilizadas para as encontrar são totalmente opostas, isto porque o *Data Mining* extrai informações que sejam implícitas, previamente desconhecidas e que sejam relevantes, ou seja, a informação dos dados utilizados é desconhecida e muito dificilmente poderá ser extraída sem o recurso a técnicas de exploração de dados. Por outro lado, o TM extrai informações que estejam explícitas no texto, ou seja, estas técnicas procuram se certificar que a informação obtida é exatamente igual àquela que foi expressa no texto inicial. [8]

As pessoas comunicam e guardam informações no formato de texto, pelo que o TM pode vir a ter um potencial muito superior ao DM, no entanto as técnicas de TM são de maior complexidade, pois analisa dados sem qualquer tipo de estrutura, com erros gramaticais, com abreviaturas e que podem estar descritos de forma desordenada [9].

Os diversos textos escritos diariamente, desde opiniões a anotações clínicas, geraram uma crescente motivação para extrair o conhecimento que, outrotora não era aproveitado, o que gerou, ao longo dos anos uma evolução das técnicas utilizadas para TM, sempre com o objetivo de exprimir o verdadeiro significado do texto, sem a necessidade de um intermediário humano [8], sendo que cada vez é mais comum verem-se aplicações e processos automatizados de TM, o que se deve essencialmente à disponibilidade de um número crescente de documentos disponíveis que incluem, por norma, informação não estruturada e/ou semiestruturada [10].

O TM é uma área em expansão que visa recolher informações úteis de textos, nos diversos idiomas existentes. Os dados não estruturados dificultam o seu aproveitamento e utilização, o que aumenta a complexidade de utilizar algoritmos, comparativamente com os dados estruturados. [8]

Com a aplicação de técnicas de TM é possível extrair informação de recursos em forma de texto, passível de ser transformada em conhecimento, através do cruzamento de técnicas como Natural Language Processing (NLP), DM, recuperação de informações, análise de texto, agrupamento, categorização, visualização e Aprendizagem Automática (AA). Através destas é possível classificar e identificar padrões a partir de diferentes tipos de documentos [11].

Para este estudo iremos utilizar a classificação de texto que é uma tarefa de TM que permite, através da definição de um conjunto de regras lógicas e treino de um modelo, classificar documentos de acordo com o conjunto de categorias que outrora seriam atribuídas manualmente por especialistas [12].

O TM aplicado ao texto biomédico é denomidado de BioNLP e consiste nos métodos e no estudo de desenvolvimento de técnicas de TM para textos e literatura dos domínios biomédico e da biologia molecular. Existem diversos desafios no BioNLP que continuam a apresentar dificuldades no desenvolvimento desta área, no entanto o facto de existir um crescimento exponencial da literatura biomédica cria uma condição motivadora para contruir melhores técnicas e métodos.

Esta dissertação aplicará técnicas e metódos de TM em registos eletrónicos médicos, com o objetivo de desenvolver estas técnicas num caso português, com dados em língua portuguesa e criar uma linha evolução no aproveitamento deste conhecimento que outrora ficaria perdido, que poderá melhor o sistema de registos eletrónicos médicos utilizado nos hospitais portugueses.

No subcapítulo seguinte, está descrita a minha motivação ao escrever e elaborar esta dissertação e a razão pela qual foi escolhida esta área de estudo e este tema.

1.3. Motivação

A crescente implementação do sistema de registos eletrônicos médicos nos Hospitais, com vista a apoiar o atendimento individual dos pacientes, está a provocar um aumento dos dados clínicos recolhidos diariamente [13]. Estes registos são preenchidos pelos médicos em texto narrativo com o objetivo de anotar as informações clínicas do paciente [13]. No entanto, o facto de não haver estrutura no texto produzido pelos médicos e o facto das informações divergirem de paciente para paciente, dificultam o aproveitamento destes dados. Sendo assim, as organizações de saúde, para superar este desafio utilizam a extração manual de informações das anotações clínicas [13], no entanto este processo é demasiado demorado e está suscetível a erros.

Segundo [14], existe uma tensão entre as necessidades dos médicos que preenchem os registos e as pessoas que reutilizam esses dados dos sistemas de informação das respetivas organizações, para procederem a análises exploratórias dos mesmos. Por um lado, os médicos valorizam a flexibilidade e a eficiência dos seus atendimentos, enquanto que as pessoas que procedem ao tratamento e análise dos dados valorizam a estrutura e a padronização [14].

Posto isto, este estudo visa encontrar o equilíbrio entre os médicos e as pessoas que estudam os dados, para que exista uma otimização do conhecimento a ser extraído dos registos médicos. Sendo que o cenário pretendido é que a extração de informação seja realizada de forma automatizada e que a classificação dos registos por tipo de doença seja efetuada de forma correta. Este cenário visa ajudar na evolução do aproveitamento das

possíveis informações contidas nos REM's e, consequentemente, visa contribuir para o crescimento destes sistemas dentro das organizações hospitalares portuguesas.

1.4. Objetivos

Um dos objetivos deste estudo é desenvolver um sistema que consegue extrair informação em Língua Portuguesa de dados não estruturados, encontrados nos registos eletrónicos médicos, de forma a conseguir classificá-los corretamente por diagnóstico (tipo de doença).

Outro dos objetivos deste estudo é examinar a eficácia da aplicação de técnicas de NLP e algoritmos de aprendizagem automática na classificação de documentos biomédicos de acordo com os diversos diagnósticos (doenças).

Os REM's foram fornecidos por um Hospital Português e contêm dados diversos referentes a um conjunto de atendimentos realizados entre janeiro de 2017 e janeiro de 2018. Para a extração de informação e, consecutiva, classificação recorremos a técnicas de NLP, TM e Aprendizagem Automática.

Este estudo procura responder às questões evidenciadas na tabela 1.1.

Tabela 1.1 – Questões

ID	Questões
1	É possível extrair informação de REM's não estruturados descritos em Língua Portuguesa?
2	É possível classificar corretamente os REM's por diagnóstico (tipo de doença)?
3	É possível extrair informação e/ ou classificar corretamente os REM por diagnóstico (tipo de doença), utilizando as técnicas de NLP para Língua Portuguesa?

1.5. Estrutura e Organização da Dissertação

O presente estudo está organizado em seis capítulos que pretendem refletir as diferentes fases desta investigação.

O primeiro capítulo introduz o tema da investigação, os objetivos, bem como uma breve descrição da estrutura do trabalho.

O segundo capítulo reflete o enquadramento teórico desta investigação relativamente à extração de conhecimento e classificação clínica, designado por Revisão da literatura.

O terceiro capítulo é dedicado ao sistema de arquitetura utilizada no processo de recolha, tratamento e classificação dos REM's, bem como os métodos de análise utilizados.

O quarto capítulo apresenta o desenvolvimento e aplicação do sistema de arquitetura definido nos dados recolhidos, bem como a explicação de todos os processos envolvidos.

O quinto capítulo apresenta a análise dos resultados finais desta dissertação e a sua respetiva avaliação.

No sexto e último capítulo apresentam-se as conclusões desta investigação bem como as limitações e trabalhos futuros.

Capítulo 2 – Revisão da Literatura

Com a adoção de sistemas de registo eletrónicos por parte das organizações hospitalares, tornou-se necessário recolher as informações dos REM's para apoiar os sistemas automatizados nos atendimentos e permitir que sejam utilizados para pesquisas e investigações clínicas [15]. Estes sistemas de tecnologias de informação são cada vez mais importantes e necessários para melhorar a qualidade do atendimento e, consequentemente, melhorar a precisão do diagnóstico do paciente, visto que a identificação das doenças é baseada em evidências [16]. Um dos maiores objetivos destes sistemas é conseguir recolher e analisar os dados clínicos, de forma regular, para originar rapidamente novas evidências clínicas [13].

Neste capítulo enunciamos e descrevemos várias técnicas de extração de informação e classificação de documentos que são utilizadas para este âmbito de investigação, bem como iremos analisar diversos casos de estudo para verificar o que tem sido desenvolvido na área biomédica, com o objetivo de posicionar este estudo na área científica.

2.1. Registos Eletrónicos Médicos

Os REM's contêm infindáveis informações clínicas, que ao serem analisados, podem resolver problemas relacionados à qualidade do atendimento e ao suporte da decisão clínica [17].

Os dados inseridos nos REM's são as anotações legais do estado de saúde do paciente durante o atendimento com os médicos [18], que podem incluir informações de diagnóstico, procedimentos realizados e/ ou resultados de tratamentos [19].

Os médicos introduzem os dados clínicos do paciente no sistema eletrónico do hospital em formato narrativo, visto que facilita a documentação da situação clínica e, caso seja necessário, facilita também a comunicação da mesma para outras equipas hospitalares [15]. Estes dados, embora tenham informações médicas efetivamente importantes e vantajosas, não estão estruturados (texto narrativo) e carecem de um complexo pré-processamento e análise, principalmente porque podem existir muitos erros gramaticais, erros de ortografia, gírias e ambiguidades semânticas [19].

De facto, os registos clínicos abrangem diversos estilos de narrativas, sendo que podem diferir de tamanho, existindo narrativas mais curtas e outras mais longas, podem estar preenchidas com muitas abreviações, acrónimos ou dialetos locais, e caso o sistema não tenha um corretor ortográfico, podem conter também muitos erros ortográficos [20]. Não esquecendo que o contexto de cada registo eletrónico, pode divergir de paciente para

paciente, de consulta para consulta e de profissional de saúde para profissional de saúde, aumento a complexidade do tratamento dos dados.

Pelo que, uma das componentes mais críticas na análise dos dados descritos nos REM´s é a extração da informação [15], que consiste na seleção dos termos mais relevantes e, consequentemente, na sua análise, tornando-se em informação e, naturalmente, em conhecimento. Esta extração de conhecimento poderá trazer como benefícios a melhoria da gestão de saúde do paciente [21], a simplificação dos processos de tomada de decisão e, consequentemente, a melhoria da prestação de assistência médica [3].

2.2. *Text Mining em Língua Portuguesa*

A linguagem de programação Python tem uma comunidade notória de utilizados, principalmente nas áreas de NLP e aprendizagem automática, especialmente na língua inglesa. No entanto, o processamento computacional da língua portuguesa nesta linguagem de programação é cada vez mais desenvolvido e as técnicas e metódos existentes são adaptadas, embora existam sempre limitações.

Atualmente, a quantidade de informações que podem ser utilizadas para extrair conhecimento em língua portuguesa é praticamente infinita. Esta linguagem de programação contem ferramentas que possibilitam extrair conhecimento e fazer o processamento NLP, tanto para língua inglesa como para língua portuguesa. Um desses exemplos é a biblioteca NLTK, que apresenta técnicas adaptadas e, inclusive, contem uma lista de stopwords na língua portuguesa.

O processamento em língua portuguesa apresenta algumas complexidades que a língua inglesa não tem, como por exemplo, a complexidade dos verbos e a acentuação.

No capítulo 2.3 está descrito os vários processos de NLP aplicados a dados biomédicos que são utilizados.

2.3. *Biomedical Natural Language Processing*

Com a adoção de sistemas de REM´s por parte das organizações hospitalares, o interesse em analisar estes dados desenvolveu uma crescente necessidade de aperfeiçoamento do processamento dos mesmos, tornando-se numa etapa crítica para conseguir obter bons resultados [22]. Devido à estrutura de texto narrativo, recorremos ao processamento dos dados, utilizando técnicas de NLP que já demonstraram ter sucesso na área médica [17].

O NLP consiste num processo computacional e automatizado para analisar texto não estruturado, como é o caso de estudo abordado nesta dissertação, para encontrar

informações e conhecimentos que, normalmente, eram difíceis de se obter com outros processos e métodos [23]. O NLP tem como objetivo conseguir representar o verdadeiro significado e a intenção do texto escrito pelo seu autor, que neste estudo são os médicos [24].

Um sistema de NLP pode incluir análises morfológicas que consistem no estudo das flexões e derivações das palavras e das suas classes, análises semânticas que estudam o significado da frase e análises sintáticas que compreendem a análise da organização das palavras nas frases. Alguns dos métodos utilizados nestes sistemas são a extração de informação (EI), reconhecimento de entidade nomeada, sumarização, entre outras.

O processamento de dados reduz significativamente o tamanho dos documentos e existem várias técnicas que podem ser utilizadas, como por exemplo, a colocação de todas as letras em minúsculas, a *tokenization* onde o texto é dividido num conjunto de *tokens*, a remoção da pontuação, a remoção das *stopwords* que consistem em palavras que aparecem com frequência e não acrescentam significado semântico à análise, o POS-TAG (*part-of-speech tagging*) que se baseia na marcação das palavras de acordo com a classe gramatical a que pertence e o *stemming* que consiste na redução das palavras à sua raiz ou radical.

Após o tratamento dos dados procede-se à extração e seleção dos atributos que consiste na identificação das palavras/ expressões mais importantes num documento. Nesta etapa pode-se utilizar diversas técnicas, tais como, o TF-IDF (*Term Frequency – Inverse Document Frequency*) que avalia a frequência de ocorrência de um termo num documento, o *Word embedding* que consiste no mapeamento das palavras em vetores, o LSA (*Latent Semantic Analysis)* que agrega palavras e documentos em vetores com contexto semelhante e o LDA (*Latent Dirichlet Allocation*) que é um modelo probabilístico em que cada documento é descrito pela distribuição dos tópicos e cada tópico é descrito pela distribuição de palavras.

O sistema de NLP biomédico é uma subárea do NLP que é dirigido para textos relativos a biologia, a medicina e/ ou a química, também conhecido por BioNLP. O texto biomédico é heterogêneo, visto que as anotações escritas pelos médicos durante os atendimentos são mais informais, podendo conter gírias e/ou abreviações, ao contrário dos artigos científicos e/ ou diretrizes de saúde pública que têm um estilo de escrita formal, o que se traduz numa necessária adequação das ferramentas de NLP a cada tipo de fonte do texto apesar de apresentarem informação da mesma área [25].

Ao longo do tempo e da evolução das técnicas computacionais de análise de dados, foram surgindo diversos sistemas de BioNLP conforme as necessidades de cada autor. Embora apenas existam sistemas para dados clínicos em língua inglesa, no seguinte capítulo existe um resumo dos mesmos, pois eles são extremamente eficazes e podemos verificar as metodologias utilizadas para depois criar um sistema de extração de conhecimento para dados língua portuguesa, que é o principal objetivo deste estudo.

2.4. Sistemas biomédicos de NLP

Neste capítulo iremos descrever os sistemas biomédicos de NLP em língua inglesa mais utilizados e os seus respetivos casos de estudo. Estes sistemas foram desenvolvidos ao longo do tempo com objetivo de ajudar e facilitar a extração de informação de relatórios médicos, sejam eles notas clínicas ou relatórios de um exame específico. Na tabela 2.2., podemos observar os casos de estudo de cada sistema, bem como o tipo de dados utilizado.

Tabela 2.2 - Casos de Estudo por Sistema de NLP Biomédico

Sistema NLP	**Aplicação**	**Tipo de Dados**	**Referência**
GATE	*Evaluation of Smoking Status*	REM	[26]
	Automating Tissue Bank Annotation	Relatórios Patológicos	[27]
HITEx	*Extracting principal diagnosis, co-morbidity and smoking status for asthma research*	REM	[28]
	Discovery Research in Rheumatoid Arthritis	REM	[29]
	Prediction of Chronic Obstructive Pulmonary Disease	REM	[30]
cTAKES	*Extracting Cancer Phenotypes*	Anotações Clínicas	[31]
	Discovering Peripheral Arterial Disease	Anotações de Radiologia	[32]
	Case Definition of Crohn's Disease and Ulcerative Colitis	REM	[33]
	Identification cardiovascular risk factors	Anotações Clínicas	[34]
	Identification of Surveillance Colonoscopy in Inflammatory Bowel Disease	REM	[35]
MedLEE	*Extracting comorbidity information*	Anotações Clínicas	[36]
	Early recognition of multiple sclerosis	REM	[37]

Sistema NLP	Aplicação	Tipo de Dados	Referência
	Identification of Findings Suspicious for Breast Cancer	Relatórios de Mamografia	[38]
	Information Retrieval in Thoracic Radiology	Relatórios de Raio-X ao Peito	[39]
MMTx	*Extracting Structured Information*	Relatórios Patológicos	[40]

2.4.1. GATE

O sistema GATE (General Architecture for Text Engineering) foi desenvolvido na Universidade de Sheffield e tem uma arquitetura de código aberto em Java, sendo utilizado para o processamento de linguagem natural, incluindo extração de informações em vários idiomas. Este sistema compreende três elementos principais o GDM (GATE Document Manager) que faz a gestão de documentos, nomeadamente, armazena todas as informações que o sistema gera ao processar os textos, o CREOLE (Collection of Reusable Objects for Language Engineering) que suporta a integração dos módulos escritos num idioma em qualquer plataforma e o GGI (GATE Graphical Interface) que avalia os módulos e apresenta ferramentas de visualização [40].

Para este sistema, analisamos dois casos de sucesso que consistem na avaliação do tabagismo através de REM [25] e na automatização de anotações no banco de tecidos através de relatórios patológicos [26].

2.4.2. HITEx

O sistema HITEx (Helth Information Text Extraction) é uma aplicação de software de NLP de código aberto em Java que foi desenvolvido com base na estrutura do sistema GATE para extrair informações específicas sobre diagnósticos e tabagismo [19], sendo que foi construído por um grupo de pesquisadores do Hospital Brigham and Women e da Harvard Medical School. Este software de NLP determina a estrutura dos registos de texto não estruturados e gera um documento anotado que identifica as variáveis mais interessantes. Este sistema tem diversos módulos, tais como o tokenizer, POS Tagger, N-gram tool, entre outros [27].

Este sistema foi criado para extrair diagnósticos de comorbidade e tabagismo dos REM [27], no entanto existem outros casos de sucesso, nomeadamente, na pesquisa de da artrite reumatoide [28] e na previsão de doença pulmonar obstrutiva crônica [29].

2.4.3. cTAKES

O sistema cTAKES (clinical Text Analysis and Knowledge Extration System) foi desenvolvido pela Clínica Mayo e é um software de NLP biomédico de código aberto [33]. Este sistema foi preparado especificamente para o domínio clínico, de forma a processar e extrair semanticamente as informações com o objetivo de diminuir a heterogeneidade dos dados da área clínica [16].

O cTAKES tem um conjunto de ferramentas de NLP para processar o texto clínico e, consequentemente, extrair informação. Algumas destas ferramentas são o tokenizer, o POS Tagger, named entities, entre outros [31].

Alguns dos casos de sucesso deste sistema consistem em extrair informação de REM, nomeadamente, extrair casos com fenótipos de cancro [30], definição da doença de Crohn e Colite Ulcerativa [32], identificação dos fatores de risco cardiovasculares [33] e identificação de colonoscopia de vigilância na doença inflamatória intestinal [34]. Este sistema também foi aplicado a anotações de radiologia para descobrir a doença arterial periférica [31].

A construção do sistema biomédico de processamento de NLP construído nesta dissertação teve por base este sistema.

2.4.4. MedLEE

Carol Friedman et al. na Universidade de Columbia desenvolveram o sistema MedLEE (Medical Language Extraction and Encoding System) para extrair informações de textos clínicos. Originalmente foi desenhado para processar anotações de radiologia, no entanto e devido ao seu desempenho o âmbito foi alargado para outros tipos de dados clínicos [41].

Devido à ampliação do campo de ação foi possível extrair informações de comorbidade de anotações clínicas [35], reconhecer precocemente a esclerose múltipla de REM [36] e identificar descobertas suspeitas do cancro da mama através de relatórios de mamografia [35].

2.4.5. MMTx

O sistema MMTx (MetaMap Transfer) é a versão Java do MetaMap que foi desenvolvido pelo US National Library of Medicine (USNLM). Este sistema é utilizado para recuperar informações, extrair informações biomédicas e, também, extrair outros tipos de informações com conceitos anatômicos ou de ligação molecular [42].

Este sistema demonstrou ser viável para a extração de informação estruturada através de relatórios patológicos [39].

2.5. Classificação de Texto Biomédico

A classificação de texto é uma tarefa supervisionada de aprendizagem automática e consiste no processo de classificar um documento numa categoria predefinida, como por exemplo, se A é um documento e {bom, mau} é um conjunto de categorias, este método irá atribuir uma das categorias do intervalo ao documento, ou seja, A foi classificado como bom estado. Também podemos ter em conta que um documento pode ter mais do que uma categoria associada (Classificação Multirótulo), no entanto para este estudo apenas será considerado que para cada documento exista apenas uma única categoria [41], que pode ser dividida em dois tipos de classificação, a binária que compreende apenas um conjunto de duas categorias, como por exemplo, “Sim” e “Não”, e a Multiclasse que pode apresentar um conjunto com mais de duas categorias. Neste último caso, a classificação torna-se mais complexa, devido ao facto de todos os algoritmos terem sido construídos para problemas de duas categorias [42].

Devemos relembrar que cada registo eletrónico médico gera milhares de palavras/ expressões, pelo que antes de utilizarmos algoritmos de classificação é necessário aplicar técnicas de processamento de linguagem natural, para selecionar e extrair os atributos mais importantes e relativos à área de medicina. Quanto melhor for o processamento e seleção dos atributos, maior a eficiência e precisão na classificação dos registos. Por outras palavras, o NLP visa entender o significado do texto como um todo, de forma a que a aprendizagem automática consiga classificar corretamente.

Tal como referido anteriormente, para termos o melhor classificador possível, teremos de lidar com certas dificuldades, nomeadamente, a classificação de texto não estruturado, pelo que a classificação deve ser feita com base em atributos, a manipulação de um grande número de atributos, pelo que se deverá recorrer a técnicas como o TF-IDF e a seleção e adequação do melhor modelo de aprendizagem automática.

Após a seleção e extração dos atributos mais importantes, divide-se o conjunto de dados em treino e teste, e aplica-se um algoritmo de aprendizagem automática ao conjunto de dados de treino para este preparar o classificador. O classificador já treinado é de seguida testado usando o conjunto de dados teste. Os algoritmos de classificação de aprendizagem automática podem ser o *Naïve Bayes*, a Árvore de Decisão, a Rede Neural, Máquinas de Vetor de Suporte ou técnicas híbridas, entre outros [43]. As métricas de avaliação dos modelos desempenham um papel crítico na adequação do melhor classificador, pois são a base de comparação entre os algoritmos. Também, é importante referir que a etapa da seleção das métricas mais adequadas a um problema é crucial. Uma das fórmulas mais utilizadas para avaliar a eficácia dos algoritmos é a *Accuracy*, que calcula a percentagem de classificações corretas tendo em conta o global das mesmas. Na etapa final, se a *Accuracy* do classificador

for aceitável, o modelo está pronto para ser utilizado para classificar novas instâncias do mesmo tipo de documentos [43].

Resumindo, as principais etapas do processo de classificação são o pré-processamento de documentos, a extração e seleção de atributos, a seleção de modelos de aprendizagem automática, o treino e teste do classificador e, por fim a avaliação do mesmo [43].

Na área de classificação biomédica, um dos grandes desafios é conseguir que estes sistemas sejam úteis para os profissionais de saúde e/ ou investigadores biomédicos, pelo que é necessário que sejam aperfeiçoados e estudados de forma a que exista uma cooperação contínua com a sociedade biomédica, de forma a garantir não só a mitigação das suas necessidades como também a evolução do TM nesta área, e consequentemente a aplicação destas ferramentas na prática. Outro desafio é quantidade de áreas e subáreas extremamente específicas dentro da Biomedicina que acaba por restringir o estabelecimento de conexões de descobertas e métodos entre as diferentes áreas de investigação [44].

Os benefícios gerados pelo estudo de algoritmos e métodos na área biomédica são, por exemplo, promover uma triagem automatizada de documentos, reduzir o trabalho humano necessário para realizar revisões sistemáticas [45], facilitar a procura de informações, descobrir padrões que não são facilmente encontrados devido ao grande volume de dados e automatizar a extração de informação de forma a criar métodos de gestão de conhecimento biomédico [44].

De seguida são apresentados os casos de estudo na área de classificação biomédica onde serão comparadas abordagens e métodos utilizados para este tipo de problema.

2.6. Casos de Estudo de Classificação de Registos Eletrónicos Médicos

Neste capítulo abordamos diversos casos de estudo relativos à classificação de registos eletrónicos médicos com o objetivo de verificar o que tem sido desenvolvido nesta área. Na tabela 2.3, podemos observar o âmbito do estudo, os métodos de processamento e extração dos atributos utilizados, como também, o sistema biomédico de NLP (caso, seja aplicável), os algoritmos utilizados para a classificação, bem como quais as métricas utilizadas para avaliar os respetivos estudos.

Tabela 2.3 - Casos de Estudo de Classificação

Aplicação	Métodos	Algoritmos	Métricas	Ref.
Automatic Infection detection	*- Text segmentation;* *- Rule-based approaches;* *- Threshold based approach.*	*- Logistic regression;* *- Naïve Bayes*; *- Gradient boost*; *- Random forest*.	*- Positive predictive value* (PPV); *- Sensitivity;* *- Area under the receiver operating characteristic curve* (AUC); *- F1-score.*	[46]
Automatic Quality of Life Prediction	*- Bag-of-concepts;* *- Bag-of-words;* *- Stopwords*; *- Correlation-based Feature Subset;* *- Information Gain.*	*- Support Vector machine (*SVM*).*	*- Positive agreement;* *- Negative agreement;* *- Kappa statistic.*	[47]
Automatic Classification of Foot Examination Findings	*- Bag-of-words;* *- Stemming.* -MMx	- SVM;	*- Kappa statistic;* *- Accuracy.*	[48]
Automatic Methods to Extract New York Heart Association	- NLP *web-based search engine* - PIER; *- Rule-based method;* *- Stopwords*; *- Lexical variant generation;* *- Bag-of-words;* *- N-gram features.*	- SVM;	*- Precision;* *- Recall;* *- F-measure.*	[49]
Discovery Research in Rheumatoid Arthritis	- HITEx.	*- Logistic regression.*	*- PPV;* *- Sensitivity;* *- Specificity.*	[29]

Aplicação	Métodos	Algoritmos	Métricas	Ref.
Surveillance Colonoscopy in Inflammatory Bowel Disease	- *Automated retrieval console* (ARC).	- *Automated retrieval console* (ARC).	- *Sensitivity;* - *F-measure.*	[35]
Early recognition of multiple sclerosis	- MedLEE.	- *Naïve Bayes.*	- *Receiver operating characteristics area under the curve* (ROC AUC); - *Sensitivity;* - *Specificity.*	[37]

Em 2017 o *National Natural Science Foundation of China* e o *Shanghai Science and Technology* desenvolveram um sistema apto para detetar automaticamente a infeção através de registos eletrónicos médicos e demonstrou ter um bom desempenho ao propor automaticamente as decisões médicas, o que poderá auxiliar os profissionais de saúde na identificação deste problema nos pacientes. Na elaboração deste sistema, recorreram a métodos de aprendizagem automática, NLP e construíram uma lista de palavras manualmente para ajudar na extração de atributos [46].

Com o objetivo de prever os resultados dos pacientes relativos à qualidade de vida relacionada à saúde, um grupo de pessoas do *National Center for Research Resources e do NIH Roadmap for Medical Research*, investigaram diversas técnicas de NLP para obter o melhor resultado sendo que para a extração dos atributos o uso da abordagem *bag-of-concepts* apresentou mais vantagens do que a técnica bag-of-words e para a seleção dos atributos o método *Information Gain* demonstrou ser melhor que o método *Correlation-based Feature Subs*et [47].

Outro caso de estudo interessante foi elaborado em 2008 e procurou utilizar técnicas de NLP e aprendizagem automática para a classificação automática de descobertas do exame do pé, através de texto não estruturado presente nos relatórios clínicos. Os resultados finais mostraram que estes métodos são favoráveis, sendo uma abordagem viável à revisão manual efetuada pelos profissionais de saúde, além de que pode melhorar a qualidade e a segurança no atendimento [48].

Em 2017, três membros da *University of* Minnesota, investigaram métodos automáticos para extrair a *New York Heart Association*, também através de métodos de NLP e aprendizagem automática. Desenvolveram um método baseado em regras e compararam

com outros dois métodos de aprendizagem automática, sendo o que apresentou melhor resultado para o objetivo definido foi o algoritmo SVM com *N-gram features* [49].

Outro estudo que demonstrou um sistema de classificação eficaz foi realizado em 2010 e contribuiu para a investigação e identificação de artrite reumatoide nos REM´s, onde evidenciaram uma nova abordagem com bons resultados utilizando os dados completos dos registos para a identificação de pacientes com esta doença [29].

Em 2012 foi desenvolvido um estudo de classificação de REM´s para identificar a colonoscopia de vigilância em pacientes com doença inflamatória, com o objetivo de testar a viabilidade e a precisão do ARC que é um software baseado em NLP que permite a classificação de documentos em texto. Este estudo conclui que o ARC é viável quando comparado à classificação manual de REM [35].

Um estudo com o objetivo de identificar precocemente a esclerose múltipla analisando os REM´s através de ferramentas de NLP, obteve bons resultados e conseguiu criar condições para futuramente ser possível eliminar o erro de diagnóstico médico ao identificar precocemente a doença [37].

Ao nível global dos estudos apresentados podemos verificar que os métodos, algoritmos e métricas são diferenciados e foram adequados a cada objetivo de análise. Relativamente aos métodos mais utilizados nos casos de estudo apresentados foram o *Bag-of-words vetor* e a remoção das *Stopwords*. Como já foi referido anteriormente a Medicina é uma área bastante heterógena e ampla em termos de conceitos e áreas de investigação e pesquisa, o que justifica os diferentes métodos de tratamento e seleção de atributos, visto que cada estudo pretendia retirar informações específicas relativa a um tipo de doença. Comparando os algoritmos utilizados, houve três que são mais evidentes nestes estudos, nomeadamente, o SVM, o *Naïve Bayes* e a Regressão Logística. Por fim, foram utilizadas diversas métricas, tais como a AUC, a *Sensitivity* e a *Specificity*. A diferenciação dos métodos, métricas e algoritmos utilizados neste casos de estudo advém da diversidade dos temas, dos inputs médicos para realizar os estudos e dos objetivos dos autores para as suas pesquisas, pelo que aumenta a complexidade da escolha das melhores metodologias para criar o sistema biomédico que é proposto nesta dissertação e dificulta a comparação com outros casos de estudo já realizados.

No subcapítulo 2.7., está descrito a definição das métricas mais utilizadas nos problemas de classificação multi-classe, bem como é apresentado casos de estudo na área médica onde foi utilizado uma abordagem de classificação multi-classe e quais as métricas para avaliação dos modelos construídos.

2.7. Métricas para Classificação Multi-Classe

Existem diversas métricas gráficas e não gráficas para avaliar algoritmos de classificação e é extremamente importante conhecer o significado de cada métrica para conseguir avaliar corretamente o resultado de cada algoritmo [50].

Nos problemas de classificação, o *dataset* de treino é usado para construir um ou vários modelos de classificação, com o objetivo de prever para cada registo a classe respectiva. Os resultados destes modelos de aprendizagem automática são avaliados para se concluir a eficiência dos mesmos e dependendo dos casos de estudo, para efetuar também uma comparação entre os diversos algoritmos aplicados [50]. Uma das tabelas mais conhecidas e utilizadas para avaliar os modelos é a matriz de confusão que mostra os reqgistos que foram corretamente classificados e os registos que foram incorretamente classificados [51]. No subcapítulo 2.7.1. está referido uma explicação da matriz de confusão.

2.7.1. Matriz de confusão

A matriz de confusão é uma tabela que regista o número de registos da classificação verdadeira versus o número de registos da classificação predita. Na tabela 2.4 está evidenciado a matriz de confusão onde podemos verificar os valores verdadeiros e os valores falsos.

Para o problema de classificação as classes são colocadas nas linhas e nas colunas "Positivo" e "Negativo", na mesma ordem. Pelo que, os registos corretamente classificados estão localizados na diagonal (diagonal central, caso existam mais de duas classes, como é o caso do nosso estudo), sendo que na parte superior desta diagonal situam-se os Falsos Positivos, que representam um erro de tipo I, ou seja, rejeitar a hipótese nula quando ela é verdadeira. Na parte inferior da diagonal situam-se os Falsos Negativos, que representam um erro do tipo II, ou seja, não rejeitar a hipótese nula quando ela é falsa.

Tabela 2.4 – Matriz de confusão

		Classificação Verdadeira	
		Positivo	Negativo
Classificação Predita	Positivo	Verdadeiro Positivo (VP)	Falso Positivo (FP)
	Negativo	Falso Negativo (FN)	Verdadeiro Negativo (VN)

2.7.2. Accuracy

Tal como referido no subcapítulo 2.5., os problemas de classificação de aprendizagem automática que tenham mais de duas classes são denominadas de classificação multi-classe. A maior parte das métricas desenvolvidas foram criadas para problemas de classificação de duas classes (binárias) [52], pelo que torna a avaliação dos modelos de aprendizagem automática mais complexa. No entanto, para a análise de um classificador multi-classe podemos utilizar uma métrica muito conhecida e utilizada, denominada de Accuracy, que mede a proporção de classificações corretas sobre o número total de classificações realizadas [52] [51].

A *Accuracy* é a probabilidade da previsão do modelo estar correta e é calculada diretamente a partir da matriz de confusão. A fórmula da *Accuracy* considera a soma dos elementos Verdadeiros Positivos e Verdadeiros Negativos no numerador e a soma de todas as entradas da matriz de confusão no denominador. Os verdadeiros positivos e os verdadeiros negativos são os elementos corretamente classificados pelo modelo e estão na diagonal principal da matriz de confusão, enquanto o denominador também considera todos os elementos fora da diagonal principal que foram classificados incorretamente pelo modelo. No caso da classificação multi-classe, a *Accuracy* calcula uma medida geral relativa à capacidade de previsão correta do modelo em todo o conjunto de dados [52].

2.7.3. Precision

A *Precision* resulta da divisão dos registos verdadeiros positivos, pelo número total de registos previstos positivamente, ou seja, a soma dos verdadeiros positivos e dos falsos positivos [52].

A *Precision* mostra a percentagem na qual podemos confiar no modelo quando este prevê os registos como positivos.

2.7.4. Recall

O *Recall* é a divisão dos registos verdadeiros positivos divididos pelo número total de registos classificados positivamente, ou seja, a soma dos verdadeiros positivos e dos falsos negativos [52]. O *Recall* mede a precisão preditiva do modelo para a classe positiva.

2.7.5. F1-Score

O F1-Score avalia o desempenho do modelo de classificação a partir da matriz de confusão, sendo que agrega as medidas de *Precision* e *Recall* sob o conceito de média harmônica.

Esta métrica resulta numa média ponderada entre *Precision* e *Recall*, onde a pontuação apresenta valores entre 0 e 1, sendo 1 o melhor valor possível, visto que significa que a contribuição relativa à *Precision* e ao *Recall* são iguais. Quando a *Precision* ou o *Recall* apresentam valores próximos a 0, o valor do *F1-Score* sofre um grande decréscimo, visto que a média harmônica tende a dar mais peso aos valores mais baixos. Podemos evidenciar que o F1-Score apenas considera a classe positiva, logo, os registos verdadeiros negativos não têm importância [52].

2.7.6. Caso de Estudo de Métricas Utilizadas

No caso de estudo [53], foi proposto uma abordagem metodológica envolvendo a integração de várias séries heterogêneas de cancro de pele e, posteriormente, a aplicação de classificador multi-classe, com o objetivo de fornecer aos médicos uma ferramenta de suporte de diagnóstico inteligente baseada no uso de um conjunto de biomarcadores selecionados, que simultaneamente distingue diferentes estados de pele relacionados ao cancro. Neste estudo foi utilizado o classificador *Support Vector Machine* (SVM) e avaliação do modelo foi efetuada com utilizando a matriz de confusão e a *accuracy* que obteve um resultado de 92%, com um total de 7 classes de diferentes estados de pele relacionados ao cancro.

Neste caso de estudo [54], foi testado o desempenho de dois algoritmos híbridos de classificação multiclasse, nomeadamente, a árvore de decisão e o *Naïve Bayes*. Para comparar estes algoritmos foram utilizadas as métricas *Accuracy, Precision e Recall*. Os resultados experimentais indicaram que os métodos propostos produziram bons resultados na classificação de problemas multi-classe, e obtiveram resultados acima de 80%.

Neste último caso de estudo [55], foi pretendido encontrar correspondência entre as imagens e as classes de diagnósticos atribuídos pelos patologistas. A avaliação dos modelos construídos, usando configurações de 5 e 14 classes, mostrou valores médios de *Precision* até 81% e 69%, respectivamente. Este estudo demonstrou que o classificador multi-classe pode realizar com sucesso a localização e a classificação de várias imagens em classes.

No capítulo 3 é abordado a arquitetura e construção do sistema biomédico que tem como base as técnicas utilizadas nos sistemas atualmente existentes para dados biomédicos em língua inglesa, bem como a explicação do algoritmo escolhido para a classificação.

Capítulo 3 - Arquitetura do Sistema Biomédico

Neste capítulo é descrito a arquitetura do sistema desenvolvido com o objetivo de classificar corretamente os REM´s em português pelos diversos diagnósticos médicos, referindo detalhadamente as diferentes etapas, nomeadamente, o processamento NLP e o algoritmo de classificação mais adequado para este problema de classificação multiclasse.

3.1. Requisitos do sistema

O sistema descrito neste estudo foi desenvolvido tendo em conta dois objetivos cruciais que estão descritos na tabela 3.5. Todos os métodos definidos na secção 3.2. tiveram o foco de cumprir com os requisitos previamente definidos.

Tabela 3.5 - Requisitos do Sistema

ID	Requisitos do Sistema
1	O sistema deve ser capaz de extrair a informação mais relevante dos REM´s em Língua Portuguesa.
2	O sistema deve ser capaz de classificar corretamente os REM's por diagnóstico (tipo de doença) em Língua Portuguesa.

3.2. Arquitetura do Sistema Biomédico

A arquitetura do sistema biomédico apresentado neste estudo visa extrair de forma correta os dados de registos eletrónicos médicos, de forma a conseguir classificar corretamente cada um num dos diagnósticos (doenças) existentes.

Na figura 3.1, está demonstrado a arquitetura macro deste sistema biomédico elaborada com base na definição dos requisitos definidos na secção 3.1.

A primeira etapa baseia-se na extração dos registos eletrónicos médicos da base de dados do Hospital para depois proceder-se ao processamento de NLP, desenvolvido especificamente para este estudo na linguagem de programação Python. Este processamento engloba diversas tarefas, nomeadamente, colocação de todas as letras em minúsculas, a *tokenization*, a remoção de *stopwords* e a pontuação, o *stemming*, o POS-TAG e o TF-IDF.

Após este processamento obtemos uma matriz de termos que com a qual podemos proceder à classificação de todos os registos por diagnóstico. Como última etapa avalia-se a performance do sistema biomédico desenvolvido e elabora-se um relatório com a referida apreciação.

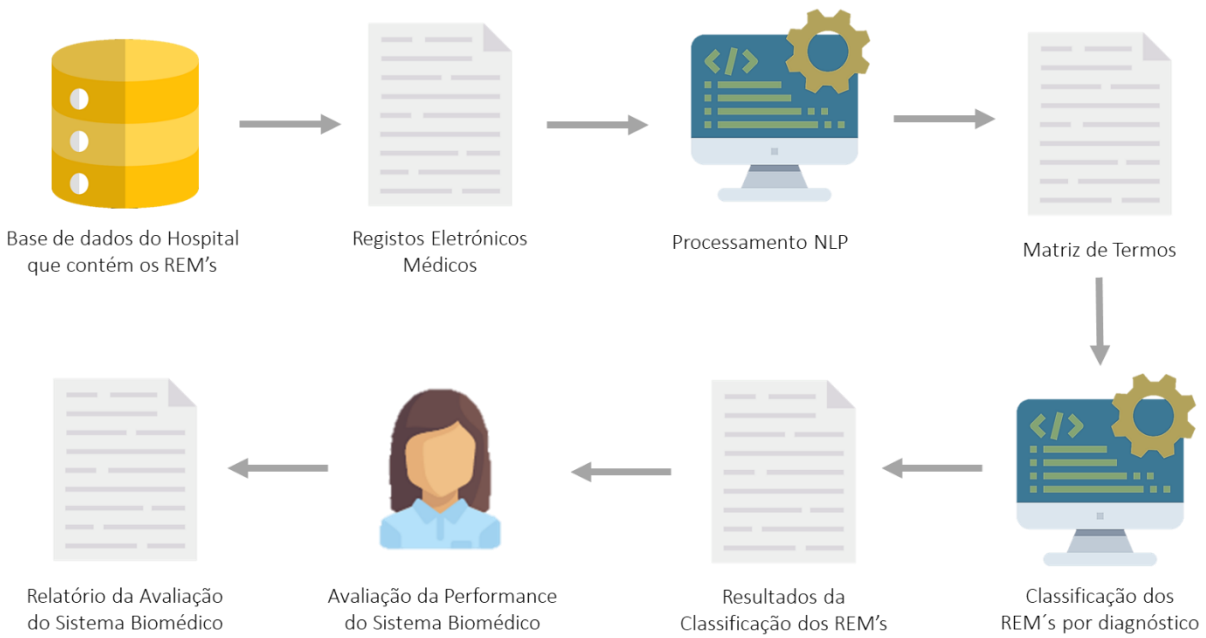

Figura 3.1 - Arquitetura do Sistema Biomédico

Nas seguintes secções encontra-se explicada detalhadamente os processos de tratamento do texto narrativo médico para conseguir extrair com sucesso as palavras mais significativas para cada diagnóstico médico.

3.2.1. Sistema NLP

O NLP consiste num processo computacional e automatizado para analisar texto não estruturado, de forma a encontrar informações e conhecimentos que evidenciam a intenção real do autor [23]. A extração de informação de texto narrativo através de outros métodos torna-se difícil e complexa [22], pelo que é escolhida esta ferramenta de tratamento de dados para o âmbito deste estudo.

Este sistema de NLP, que foi desenvolvido integralmente na linguagem de programação Python, com diversos componentes de tratamento de dados linguísticos para processar as narrativas escritas pelos médicos. Cada componente tem um objetivo diferente, no entanto no seu conjunto contribuem apenas para um propósito, o de extrair as palavras mais relevantes

de forma a classificar corretamente cada registo. Na figura 3.2, podemos verificar a sequência dos diferentes componentes referentes ao processamento dos dados desenvolvido.

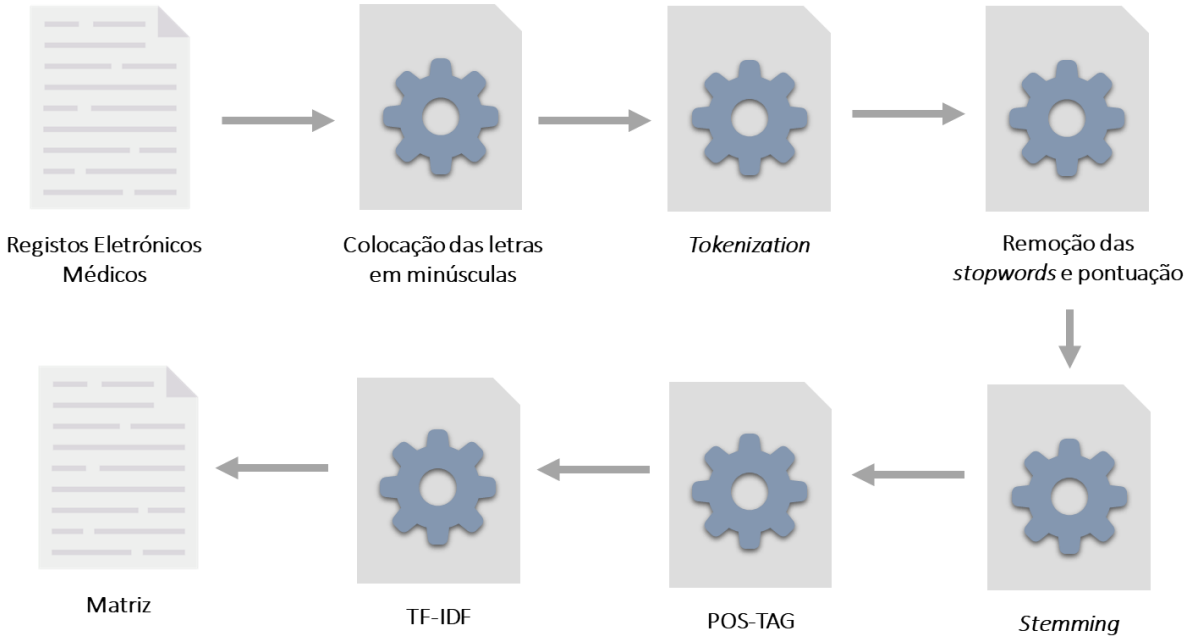

Figura 3.2 - Arquitetura do sistema NLP

Para este sistema de NLP foram escolhidas cinco componentes de limpeza e tratamento de dados e um componente de seleção e extração de termos. As primeiras cinco componentes consistem em normalizar os dados e remover aqueles que não são relevantes e a última componente serve para selecionar e extrair, do conjunto de termos que sobraram das primeiras cinco tarefas, os atributos mais relevantes e significativos de cada REM.

A primeira etapa baseia-se na colocação de todas as palavras em minúsculas, visto que o processamento NLP ao longo dos REM´s é sensível às letras maiúsculas e minúsculas e pode inferir em erro palavras que sejam idênticas, como por exemplo, “CROHN” ou “crohn”, ambas significam o mesmo, mas o sistema pode interpretá-las como diferentes.

Na segunda fase efetuamos a divisão das frases em palavras e pontuação, denominados de *tokens*. Esta fase torna-se crucial porque estes *tokens* são usados para criar um vocabulário para a etapa de extração e seleção dos atributos. De seguida foram removidas as pontuações e as *stopwords* (palavras muito frequentes que não acrescentam significado à frase), de forma a diminuir a quantidade de *tokens* realmente necessários para as próximas fases do sistema de NLP.

A biblioteca NLTK da linguagem de programação Python apresenta uma lista pré-definida de stopwords. A esta lista também foram adicionados alguns termos que eram recorrentes

nos REM's. Podemos observar na tabela 3.6 que quais as stopwords incluídas nesta lista e que serão removidas dos dados dos registo eletrónicos médicos.

Tabela 3.6 – Lista de Stopwords

Lista de Stopwords
['de', 'a', 'o', 'que', 'e', 'é', 'do', 'da', 'em', 'um', 'para', 'com', 'não', 'uma', 'os', 'no', 'se', 'na', 'por', 'mais', 'as', 'dos', 'como', 'mas', 'ao', 'ele', 'das', 'à', 'seu', 'sua', 'ou', 'quando', 'muito', 'nos', 'já', 'eu', 'também', 'só', 'pelo', 'pela', 'até', 'isso', 'ela', 'entre', 'depois', 'sem', 'mesmo', 'aos', 'seus', 'quem', 'nas', 'me', 'esse', 'eles', 'você', 'essa', 'num', 'nem', 'suas', 'meu', 'às', 'minha', 'numa', 'pelos', 'elas', 'qual', 'nós', 'lhe', 'deles', 'essas', 'esses', 'pelas', 'este', 'dele', 'tu', 'te', 'vocês', 'vos', 'lhes', 'meus', 'minhas', 'teu', 'tua', 'teus', 'tuas', 'nosso', 'nossa', 'nossos', 'nossas', 'dela', 'delas', 'esta', 'estes', 'estas', 'aquele', 'aquela', 'aqueles', 'aquelas', 'isto', 'aquilo', 'estou', 'está', 'estamos', 'estão', 'estive', 'esteve', 'estivemos', 'estiveram', 'estava', 'estávamos', 'estavam', 'estivera', 'estivéramos', 'esteja', 'estejamos', 'estejam', 'estivesse', 'estivéssemos', 'estivessem', 'estiver', 'estivermos', 'estiverem', 'hei', 'há', 'havemos', 'hão', 'houve', 'houvemos', 'houveram', 'houvera', 'houvéramos', 'haja', 'hajamos', 'hajam', 'houvesse', 'houvéssemos', 'houvessem', 'houver', 'houvermos', 'houverem', 'houverei', 'houverá', 'houveremos', 'houverão', 'houveria', 'houveríamos', 'houveriam', 'sou', 'somos', 'são', 'era', 'éramos', 'eram', 'fui', 'foi', 'fomos', 'foram', 'fora', 'fôramos', 'seja', 'sejamos', 'sejam', 'fosse', 'fôssemos', 'fossem', 'for', 'formos', 'forem', 'serei', 'será', 'seremos', 'serão', 'seria', 'seríamos', 'seriam', 'tenho', 'tem', 'temos', 'tém', 'tinha', 'tínhamos', 'tinham', 'tive', 'teve', 'tivemos', 'tiveram', 'tivera', 'tivéramos', 'tenha', 'tenhamos', 'tenham', 'tivesse', 'tivéssemos', 'tivessem', 'tiver', 'tivermos', 'tiverem', 'terei', 'terá', 'teremos', 'terão', 'teria', 'teríamos', 'teriam', 'u', 'k', 'l', 'q', 'w', 'e', 'r', 's', 't', 'u', 'i', 'o', 'p', 'a', 'd', 'f', 'g', 'h', 'i', 'j', 'l', 'm', 'z', 'x', 'c', 'v', 'b', 'n', 'pnrp', 'inf', 'es', 'ml']

A lista de pontuações removida no sistema biomédico desenvolvido encontra-se expressa na tabela 3.7.

Tabela 3.7 – Lista de pontuações

Descrição	Pontuações
Hífen	-
Ponto final	.
Ponto de interrogação	?
Ponto de exclamação	!
Ponto e vírgula	;
Dois pontos	:
Aspas	“
Apóstrofe	'
Vírgula	,

Na quarta etapa foi realizado o *Stemming* que consiste na redução da palavra ao seu radical, retirando os prefixos e sufixos. Este processo ajuda a retirar o excesso de dados, sintetizando palavras que têm o mesmo significado, como por exemplo, “vital” e “vitais” ficará reduzido a uma palavra “vita”.

Na última fase de limpeza e tratamento de dados foi aplicado o POS-TAG que consiste em aplicar a cada termo a sua classe gramatical, como por exemplo, nomes, verbos, adjetivos, entre outros.

Os processamentos de limpeza e tratamento dos REM’s foram realizados através da ferramenta de NLTK (*Natural Language Toolkit*).

Relativamente à etapa de extração e seleção de atributos foi escolhido o TF-IDF visto que esta abordagem não efetua apenas a contabilização dos termos, mas também realiza a contabilização da quantidade de registos em que cada termo aparece.

Após a definição do Sistema NLP procedemos à definição da classificação mais adequada para o âmbito deste estudo.

3.2.2. Classificação

A classificação de texto é uma tarefa supervisionada de aprendizagem automática e consiste no processo de classificar um documento numa categoria predefinida. Existem vários tipos de classificação, no entanto para este estudo aplica-se a classificação multiclasse que consiste

num problema que pode apresentar mais de duas categorias/ classes, sendo que cada variável apenas pode corresponder a uma dessas categorias. No âmbito desta investigação existem diversos diagnósticos (categorias) e cada REM irá corresponder a um único diagnóstico. Este tipo de classificação torna-se mais complexa, devido ao facto de existirem diversas classes possíveis para cada REM, ao invés do problema de classificação binário que apenas compreende duas classes.

Naïve Bayes é um algoritmo de aprendizagem automática que apresenta um bom desempenho para problemas de classificação [56]. Este algoritmo tem várias versões, sendo que uma delas é o *Naïve Bayes* Bernoulli e outra é o *Naïve Bayes* Multinomial. Na versão *Naïve Bayes Bernoulli*, o documento é representado como um vetor binário de ocorrências de palavras, ou seja, se a palavra ocorrer no documento é atribuído o valor de 1 (um) e no caso contrário é atribuído o valor de 0 (zero) [56]. Na versão *Naïve Bayes* Multinomial, cada documento é visto como uma coleção de palavras e a ordem das mesmas é considerada irrelevante [57], sendo que o documento é representado como um vetor de contagem de palavras, ou seja, é realizado o apuramento do número de ocorrências de uma palavra nesse mesmo documento. A probabilidade de um vetor do documento é realizada por uma distribuição multinomial [56].

Para este estudo foi escolhida a versão *Naïve Bayes* Multinomial, visto que diversos estudos efetuados anteriormente descobriram que esta versão, normalmente apresenta melhor desempenho no que respeita a problemas de classificação [56] [57]. Para além disso, foi efetuado um estudo às três vertentes do *Naïve Bayes*, que se compreendem na Guassiana, no *Bernoulli* e no Multinomial, sendo que o modelo que obteve melhores resultados de *Accuracy* foi o *Naïve Bayes* Multinomial, como podemos observar no capítulo 5.

No quarto capítulo é abordado o desenvolvimento do sistema biomédico para classificação automática dos REM's, bem como é explicada a aplicação de cada processo do sistema no conjunto de dados médicos.

Capítulo 4 - Sistema Biomédico

Neste capítulo será descrito o pré-processamento realizado aos REM's, o processamento NLP às narrativas clínicas, bem como a aplicação do algoritmo de classificação aos dados clínicos finais.

4.1. Pré-Processamento dos REM's

Os registos eletrónicos médicos fornecidos pelo Hospital foram realizados entre janeiro de 2017 e janeiro de 2018, sendo que estes foram extraídos de uma base de dados do Hospital para um ficheiro Excel. Os REM contêm informação clínica descrita pelo médico durante um atendimento com um paciente. Cada registo tem diferentes áreas de preenchimento na base de dados do Hospital, nomeadamente, número de episódio clínico, data, código de especialidade, especialidade, código de diagnóstico, diagnóstico e uma narrativa sobre o estado clínico do paciente, sendo que as narrativas são preenchidas pelos médicos na altura do atendimento. No fim da consulta, os REM ficam guardados na base de dados do Hospital, podendo, mais tarde, serem extraídos. No ficheiro Excel, as colunas contêm toda a informação anteriormente descrita e cada linha corresponde a um registo diferente. Na tabela 4.8, está apresentado um exemplo de um REM em Língua Portuguesa.

A variável "Episódio Clínico" é referente ao número da consulta, sendo um identificador único do REM.

A variável "Especialidade", tal como o nome indica, refere a especialidade médica que será abordada na consulta, sendo que está diretamente ligada à variável "Código de Especialidade".

A variável "Diagnóstico" menciona a descrição do tipo de doença que foi analisada durante o atendimento, sendo que está diretamente ligada à variável "Código de Diagnóstico". O diagnóstico é introduzido posteriormente à consulta, de forma manual.

A variável "Data" representa a data em que o atendimento foi realizado.

Na tabela 4.8, podemos verificar que a variável "Diário" é referente ao estado clínico do paciente e encontra-se em texto não estruturado. Todas as narrativas clínicas divergem de consulta para consulta e não existe nenhum requisito mínimo que deve ser descrito pelo médico nestes diários, podendo ser mencionados diferentes doenças, sintomas, análises, medicamentos, informações sobre o paciente, entre outros. Neste ponto surge a dificuldade de utilizar estes dados para técnicas de previsão ou para informação estatística, pelo que emerge a necessidade de conseguir retirar informação dos mesmos.

Tabela 4.8 - Exemplo de um REM

Registo Eletrónico Médico
Episódio clínico: 17011
Especialidade: IMUNOHEMOTERAPIA
Código de Especialidade: 40730
Diagnóstico: ARTRITE REUMATOIDE
Código de Diagnóstico: 7140
Data: 06/11/17
Diário: Doente sem queixas que impeçam a administração de 280 mg de Tocilizumab EV. LOTE B3010H12+B3015H04. Peso: 50.6 Kg Colhido sangue para análises + urina tipo II. 09:00 - 10cc/h, TA 117/54 mm Hg; FC 78 ppm 09:15 - 130cc/h 10:00 - TA 140/61 mm Hg; FC 80 ppm. Terminou perfusão sem intercorrências.

Antes de efetuar a selação das variáveis mais relevantes para este estudo, foram analisados e corrigidos os erros gramaticais de todos os registos, diretamente no ficheiro Excel disponibilizado pelo Hospital Português.

Em cada registo existem sete variáveis, no entanto para este estudo não foram selecionadas todas elas. O “Episódio Clínico” não acrescenta nenhum conhecimento que possa vir a ser útil para ajudar o sistema biomédico a classificar, pois é apenas um identificador único de cada registo, pelo que foi excluído. O mesmo aplica-se às variáveis “Especialidade” e “Código de Especialidade” que não apresentam nenhuma relevância para o estudo, pois o sistema pretende classificar cada registo por diagnóstico e não por especialidade e por isso não foram selecionadas para a continuação desta dissertação. A variável “Data” não apresenta nenhuma informação que seja útil para o algoritmo, visto que esta apenas indica o dia em que a consulta ocorreu o que não permite distinguir cada registo por doença, por isso também foi excluída. O “Código de Diagnóstico” e “Diagnóstico” estão

interligados e significando o mesmo, ou seja, o código é o identificador para cada diagnóstico. Neste estudo apenas foi selecionada a variável "Diagnóstico" visto que esta permite uma maior rapidez na identificação da doença ao contrário do código. Esta variável contem as classes nas quais o algoritmo desenvolvido classifica cada registo. Por fim, a variável "Diário" contem as anotações clínicas e encontra-se em formato de texto, o que permite através de técnicas de TM, extrair os atributos mais relevantes para obter o melhor classificador possível.

Após a seleção das variáveis, foram removidos os valores omissos, sobrando para este estudo um total de 9.487 registos, sendo que este processo foi desenvolvido na linguagem de programação Python.

Após as atividades de seleção e limpeza do *dataset*, os registos estão prontos para a fase seguinte, onde é aplicado o processamento NLP, tal como explicado no subcapítulo 4.2.

4.2. Processamento das narrativas clínicas presentes nos REM's

De forma a iniciar o processamento das componentes de NLP, selecionámos apenas a coluna correspondente à narrativa clínica, pois esta encontra-se em formato não estruturado pelo que é necessário aplicar um conjunto de diversas técnicas de TM, para reduzir o texto não estruturado a um conjunto de atributos relevantes para este estudo. O sistem ade processamento de NLP foi desenvolvido integralmente na linguagem de programação Python.

Com o objetivo de conseguir uma melhor explicação do processamento realizado aos dados foi escolhido aleatoriamente uma narrativa de um registo eletrónico médico para servir como exemplo ao longo das descrições dos componentes, que está evidenciado na tabela 4.9.

Tabela 4.9 - Exemplo de REM

"Doente leva 4 sacos para urostomia para troca no domicílio."

Em primeiro lugar, o texto narrativo é colocado em letras minúsculas, utilizando uma função com a técnica *Python String lower()*. Depois deste processo, a componente do Python *nltk.word_tokenize* irá dividir o texto em diferentes *tokens* e remover a pontuação, ou seja, irá dividir a frase pelas suas palavras e pontuação, caso seja aplicável, tal como podemos observar na tabela 4.10, utilizando o exemplo suprarreferido.

Tabela 4.10 - Exemplo de *tokenization* no REM

doente	leva	4	sacos	para	urostomia	para	troca	no	domicílio

Depois da divisão por *tokens*, são removidas as *stopwords* que consistem em palavras que não acrescentam nenhum significado à frase e por isso devem ser retiradas, como por exemplo, “e”, “como”, “para”, entre muitos outros termos, utilizando uma lista já criada de *stopwords* na língua portuguesa através da função *stopwords.words('portuguese')* do Python. O quarto componente denomina-se de *stemming* e consiste em reduzir as palavras à sua forma radical, removendo os prefixos e sufixos das mesmas, para este passo foi utilizado a função *PorterStemmer().stem* do Python. Na tabela 4.11 podemos evidenciar o efeito destas componentes.

Tabela 4.11 - Exemplo de *stemming* no REM

doen	leva	4	saco	urostomia	troca	domicílio

De seguida é aplicado o processamento POS-TAG que reside em etiquetar cada *token* da frase na sua classe gramatical, como por exemplo, nomes, verbos, adjetivos, advérbios, entre outros, utilizando a função *nltk.pos_tag* do Pyhton . Nesta componente, com o objetivo de reduzir a quantidade de palavras não relevantes além das *stopwords*, foi efetuada uma seleção apenas das palavras etiquetadas como nomes próprios e nomes comuns no singular e no plural. Podemos verificar a sua aplicação na frase na tabela 4.12.

Tabela 4.12 - Exemplo de processamento POS-TAG no REM

Nome Singular	**Nome Singular**	**Nome Próprio Singular**	**Nome Singular**
doen	saco	urostomia	domicílio

Após o tratamento dos dados procede-se à extração e seleção dos atributos que consiste na identificação das palavras/ expressões mais significativas num registo. Nesta etapa foi utilizada a técnica TF-IDF, em detrimento de outros possíveis modelos existentes que se baseiam em frequências de termo absoluto, o que pode levar a que termos que ocorrem com frequência nos documentos possam tender a ‘ofuscar’ outros termos no conjunto de recursos que podem vir a ser valiosos. Desta forma o modelo TF-IDF tenta ultrapassar esta questão,

recorrendo a um fator de computação escalar ou normalizado onde a frequência dos termos (tf) é subtraída pela frequência dos documentos inversa (idf). Para este estudo foram escolhidos 1.000 (mil) atributos, segundo a componente TF-IDF.

No final deste processamento obtemos uma matriz de dimensão 9.487 registos x 1.000 atributos, que contém as palavras mais relevantes de cada registo que servem para ajudar a classificar corretamente cada REM.

4.3. Classificação

A classificação dos REM's será dividida em quatro tipos de diagnósticos que estão evidenciados na tabela 4.13 e na última coluna desta tabela é apresentado a divisão da totalidade de registos eletrónicos médicos utilizados no âmbito deste estudo. Existem 9.487 registos, no entanto para treinar e testar o classificador, dividiu-se o dataset inicial em treino e teste efetuando uma divisão de 80% e 20%, respectivamente. No dataset treino ficámos com um total de 7.589 registos e no dataset de teste sobrou um total de 1.898 registos.

Tabela 4.13 - Tipos de Diagnóstico

ID	Diagnóstico (nomenclatura utilizada pelo Hospital)	Nº de registos eletrónicos médicos
1	DOENCAS DO SANGUE E ORGAOS HEMATOP.	1727
2	DOENCAS DO SISTEMA OSTEOART. E TECIDOS CONJUNT.	1566
3	DOENCAS INFECCIOSAS E PARASITARIAS	1188
4	TUMORES (NEOPLASMAS)	3108
Total de registos no dataset treino		7.589

O objetivo desta investigação é classificar corretamente cada REM por um destes diagnósticos e para isso é utilizado o algoritmo *Naïve Bayes* Multinomial que demonstrou obter bons resultados de desempenho em problemas de classificação nestes casos de estudo [56] [57] e obteve o melhor resultado relativamente aos outros dois modelos *Naïve Bayes*, como podemos observar no capítulo 5.

Como fase final do processamento do sistema biomédico desenvolvido, calculou-se a sua eficiência consoante as métricas descritas no capítulo 5.

Capítulo 5 - Avaliação

Posteriormente ao processamento do sistema de NLP aos registos eletrónicos médicos, procedeu-se à divisão do conjunto de dados, em treino e teste, sendo que foi utilizada uma proporção de 80%/ 20%, respetivamente, para a verificação se os diagnósticos estão a ser corretamente preditos pelo sistema biomédico desenvolvido. Esta proporção foi escolhida devido ao princípio de Pareto que mostra que 80% dos efeitos vêm de 20% das causas [58].

A fase de avaliação desempenha um papel crítico na obtenção do melhor sistema de classificação possível, sendo que a seleção das métricas para calcular a sua eficiência é de extrema importância, visto que permitem a comparação entre as várias tentativas efetuadas no decorrer do aperfeiçoamento do sistema.

A maior parte das métricas desenvolvidas foram criadas para problemas de classificação de duas classes (binárias) [52], pelo que torna mais complexa a análise da eficiência do nosso sistema biomédico. No entanto, para a análise global do sistema biomédico, existe uma métrica que normalmente é utilizada em quase todos os estudos para aferir o seu desempenho, denominada de *Accuracy,* que mede a proporção de classificações corretas sobre o número total de classificações realizadas [52]. A fórmula desta métrica está evidenciada na equação 1.

$$Accuracy = \frac{\text{Nº de registos corretamente classificados}}{\text{Nº de registos classificados}} \quad (1)$$

A métrica *Accuracy* atribui um valor de precisão ao sistema comparando o diagnóstico que foi atribuído pelo médico na altura do atendimento (classe real/ verdadeira) e o diagnóstico que o classificador previu para determinado REM (classe predita).

5.1. Algoritmo Escolhido

Para escolher qual o algoritmo que mais se adequa ao sistema biomédico desenvolvido foi aplicada a fórmula da *Accuracy* aos três modelos Naïve Bayes, como método de comparação para observar qual deles é que teria melhor resultado. Para efetuar esta comparação foi escolhida esta métrica porque ela consegue avaliar o sistema de classificação, na globalidade, ao contrário das outras fórmulas.

Tabela 5.14 – Melhor algoritmo de classificação

Algoritmo de Classificação	***Accuracy***
Naïve Bayes Guassiana	85,98%
Naïve Bayes Bernoulli	88,78%
Naïve Bayes Multinomial	89,98%

Como podemos observar na tabela 5.14, os três algoritmos apresentaram bons resultados de precisão, no entanto aquele que obteve o melhor valor foi o *Naïve Bayes* Multinomial e por isso foi o escolhido para integrar o sistema biomédico desenvolvido nesta dissertação.

5.2. Avaliação do Sistema Biomédico

Após a escolha do melhor algoritmo que apresentou uma Accuracy de, aproximadamente, 90%. Isto significa que o sistema biomédico desenvolvido apenas classificou erroneamente 10% dos REM's do dataset de teste.

Na tabela 5.15 está representada a matriz de confusão que evidencia os valores na diagonal que foram corretamente classificados.

Tabela 5.15 - Matriz de confusão

		Predição			
	ID Diagnóstico	1	2	3	4
Real	1	**299**	5	20	79
	2	5	**376**	0	18
	3	5	1	**285**	7
	4	70	10	1	**717**

Podemos observar também na tabela 5.15 que as ligações entre os diagnósticos 1 e 4, que correspondem respetivamente a "DOENCAS DO SANGUE E ORGAOS HEMATOP." e "TUMORES (NEOPLASMAS)", são as que apresentam maior número de classificações

erróneas, devendo existir uma ligeira semelhança nos termos que representam estes dois diagnósticos.

Para conseguir uma análise mais detalhada do desempenho deste estudo, foram escolhidas mais três métricas, nomeadamente, a *Precison*, o *Recall* e o *F1 Score*. Estas métricas apenas são calculadas por cada classe/ categoria, que no âmbito deste estudo são os diagnósticos, pelo que apesar de não conseguirmos obter estas métricas referentes à globalidade do sistema, podemos efetuar uma comparação entre as diferentes classes. Foram escolhidas estas métricas pois foram utilizadas em casos de estudo biomédicos de classificação multi-classe para avaliar a performance do modelo construído. Para além disso, estas métrcias focam-se nos registos que foram corretamente classificados, que é o verdadeiro objetivo desta dissertação, ou seja, perceber se a classificação por registo está a ser feita corretamente ou incorretamente [52].

A métrica *Precision* é utilizada para medir a proporção do número de REM's que foram corretamente previstos relativos a uma classe, em relação ao número total de REM's que foram previstos relativamente a uma classe [52].

$$Precison = \frac{\text{Nº de registos corretamente classificados de uma classe}}{\text{Nº de registos previstos de uma classe}} \quad (2)$$

A métrica *Recall* é utilizada para medir a proporção do número de REM's que foram corretamente previstos relativos a uma classe, em relação ao número total de REM's reais dessa classe [52].

$$Recall = \frac{\text{Nº de registos corretamente classificados de uma classe}}{\text{Nº de registos reais de uma classe}} \quad (3)$$

A métrica *F1 Score* representa a média harmônica entre os valores das métricas *Recall* e *Precision*, ou seja, se o resultado for igual a 1 (um) indica que a *Recall* e a *Precision* apresentam valores perfeitos [52].

$$F1\ Score = 2\ x\ \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

Na tabela 5.16, podemos verificar as três métricas supramencionadas por cada diagnóstico e na última coluna podemos observar o número de REM's do *dataset* de teste divididos por classe.

Tabela 5.16 – Métricas por diagnóstico (classe)

ID Diagnóstico	*Precision*	*Recall*	*F1-Score*	Nº de REM's
1	80,9%	77,9%	79,4%	403
2	96,3%	96,5%	96,4%	399
3	93,5%	97,0%	95,2%	298
4	89,9%	90,2%	90,1%	798

Os números de REM's apresentados na tabela 5.16 corresponde aos valores de registos por cada classe no dataset teste.

Relativamente ao diagnóstico nº 1 "DOENCAS DO SANGUE E ORGAOS HEMATOP." podemos analisar que foi a classe que obteve os resultados mais baixos nas três métricas e foi a única que não conseguiu superar os 81%. No diagnóstico nº 2 "DOENCAS DO SISTEMA OSTEOART. E TECIDOS CONJUNT." os valores das métricas são praticamente iguais e rondam os 96%, o que demonstra uma consistência do desempenho do classificador nesta classe. Este diagnóstico apresenta o melhor resultado para a *Precision* e para o F1 Score. No terceiro diagnóstico que corresponde a "DOENCAS INFECCIOSAS E PARASITARIAS" verificamos que é o que apresenta o melhor valor para o *Recall*, comparativamente às outras classes. No quarto diagnóstico "TUMORES (NEOPLASMAS)", apesar de ser o que apresenta maior número de REM's tanto no *dataset* treino como no *dataset* teste, não é o que apresenta nem os melhores nem os piores resultados para nenhuma das métricas, tendo uma consistência de resultados que ronda os 90%, em média.

Podemos concluir que de forma geral o sistema biomédico desenvolvido neste estudo com técnicas de NLP e de Aprendizagem Automática é viável para classificar REM's pelo tipo de diagnóstico.

Capítulo 6 - Conclusões

Este estudo visou encontrar o equilíbrio entre os médicos e as pessoas que estudam os dados, para que exista uma otimização do conhecimento a ser extraído dos registos eletrónicos médicos e um maior aproveitamento de conhecimento para os profissionais de saúde. Com os registos eletrónicos médicos fornecidos pelo hospital português iniciou um processo de pesquisa para encontrar este equilíbrio e inciar um projeto que implique a utilização dos dados brutos em língua portuguesa e que não houvesse necessidade de nenhuma tradução, pois por melhor que seja o rácio de tradução, existe sempre uma percentagem de informação que fica perdida. O estado de arte evidenciado no capítulo 2, permite ter uma visão do que está a ser realizado na área médica com os registos eletrónicos médicos, quais os sistemas utilizados para extrair informação, os algoritmos escolhidos, as métricas usadas para avaliar todos estes sistemas. Também, evidenciou o que já foi atualmente desenvolvido na área de Text Mining para Língua Portuguesa. Esta dissertação muito dificilmente pode ser comparada com algum outro caso de estudo, pois é extremamente especifica, ou seja, não existe nenhum caso de estudo proponha classificar cada registo eletrónico médico por diagnóstico em português.

Os registos eletrónicos fornecidos apresentavam sete variáveis, mas nem todas foram necessárias para estudo. O "Episódio Clínico", a "Especialidade", o "Código de Especialidade", a "Data" e o "Código de Diagnóstico" não acrescentavam nenhum significa que pudesse a vir a ser útil para ajudar o sistema biomédico a classificar corretamente e por este motivo foram excluídas. Sobrando para este estudo apenas duas varíaveis, a variável "Diagnóstico" visto que esta contem as classes nas quais o algoritmo desenvolvido classifica cada registo e a variável "Diário" que contem as anotações clínicas em formato de texto, o que possibilita a extração dos atributos mais relevantes para obter o melhor classificador possível.

Para construir o sistema foi necessário definir os seus requisitos que se basearam na capacidade de extração da informação mais relevante dos registos e a capacidade de classificar corretamente os registos por diagnóstico. Para o primeiro requisito foram efetuadas diversas tentativas e aplicação de diversas técnicas de processamento, no entanto a melhor combinação surgiu da aplicação de letras minúsculas, tokanização, remoção de stopwords, remoção de pontuação, Stemming, POS-Tag e por fim a seleção dos atributos com o TF-IDF. O tratamento dos dados foi realizado exclusivamente na língua de programação Python, utilizando a biblioteca NLTK. Este processamento permitiu extrair uma grande quantidade de atributos relevantes que possibilitaram obter os resultados evidenciados no capítulo 5. Este requisito do sistema foi fundamental para criar condições de a forma a que se conseguisse

corresponder com o segundo requisito de classificar corretamente cada registo, pela classe de diagnóstico.

Para classificar a matriz de atributos em cada classe de diagnóstico, foram analisados três vertentes do algoritmo Naïve Bayes, denominadas de Naïve Bayes Guassiana, Naïve Bayes Bernoulli e Naïve Bayes Multinomial. Para comparar e escolher o melhor algoritmo foi considerado apenas a métrica *Accuracy* pois é a única que possibilita a avaliação do sistema com os quatro diagnósticos (classes). O Naïve Bayes Multinomial obteve o melhor resultado, embora nenhum dos três tivesse apresentado um resultado relativamente desinteressante. A classificação e construção dos algoritmos foi realizada exclusivamente na linguagem de programação Python, recorrendo à biblioteca Scikitlearn.

A classificação nesta dissertação não consistia em apenas duas classes, mas sim em quatro classes, “doencas do sangue e orgaos hematop.”, “doencas do sistema osteoart. e tecidos conjunt.”, “doencas infecciosas e parasitarias” e “tumores (neoplasmas)”, o que torna mais complexa a obtenção de melhores resultados, em comparação com a classificação binária. No entanto e apesar desta dificuldade, foi possível obter uma taxa de *accuracy* de 90%, referente à globalidade das classes preditas, no algoritmo Naïve Bayes Multinomial. Esta taxa demonstra que apenas 10% dos registos foram incorretamente classificados. Comparativamente aos casos de estudo de classificação multi-classe evidenciados no capítulo 2.7.6., verificamos que os valores desta dissertação não são contrários aos estudos que já foram desenvolvidos. No entanto é preciso ter sempre em conta que os temas, embora sejam todos da área médica, são diferentes e procuram resolver problemas diferentes, pelo que a comparação tem de ser sempre relativa.

Através da matriz de confusão, além de conseguir obter com a mesma todas as métricas de avaliação deste sistema, ela mostrou-nos que existe alguma parecença entre o diagnóstico 1 e o 4 que correspondem respetivamente a “doencas do sangue e orgaos hematop.” e “tumores (neoplasmas)”.

Relativamente aos resultados obtidos por cada classe, conseguimos verificar que o diagnóstico nº 1 “doencas do sangue e orgaos hematop.” foi a classe que obteve os resultados mais baixos nas três métricas, não conseguindo ultrapassar o valor 81% em nenhuma das métricas. No diagnóstico nº 2 “doencas do sistema osteoart. e tecidos conjunt.” os valores das métricas apresentam consistência e rondam os 96%, sendo que apresenta o melhor resultado na métrica *Precision e F1-score*. No terceiro diagnóstico, “doencas infecciosas e parasitarias”, verificamos que apresenta o melhor valor para o *recall,* comparativamente às outras classes. No quarto diagnóstico “tumores (neoplasmas)”, apesar de ser o que apresenta maior número de REM’S, o que se traduz num maior número de dados, não é o que apresenta nem os melhores nem os piores resultados para nenhuma das métricas, tendo uma consistência de resultados que ronda os 90%, em média.

Um dos objetivos e requisitos do sistema biomédico desenvolvido era conseguir extrair informação dos REM's, que continham dados não estruturados descritos em Língua Portuguesa. Este objetivo foi atinigido, pelo que a questão nº 1 da capítulo 1 também foi respondida, ora vejamos, o classificador apenas consegue classificar corretamente caso os dados de entrada estejam limpos e sintetizados, de forma a agrupar os atributos mais relevantes. Posto isto, é possível avaliar se o processamento dos dados foi realizado corretamente, analisando a capacidade classificativa do modelo. Pelo que podemos concluir, através da observação dos resultados apresentados no subcapítulo 1.4. serem todos acima de 77%, que o processamento dos dados conseguiu extrair informação sintetizada e relevante, dando ao classificador bons argumentos para efetuar uma atribuição correta de uma classe a cada registo.

Relativamente à questão de ser possível classificar corretamente os REM's por diagnóstico (tipo de doença), podemos considerar que se encontra respondida, visto que na globalidade do sistema, apenas 10% dos REM's, foram classificados erroneamente, o que apresenta uma taxa muito positiva relativa à capacidade de classificação.

A terceira e última questão baseia-se em avaliar a possibilidade de extração de informação e/ ou classificar corretamente os REM por diagnóstico (tipo de doença), utilizando as técnicas de NLP para Língua Portuguesa. Esta questão ficou respondida porque a língua portuguesa não apresentou dificuldades no desenvolvimento das técnicas de texto mining, visto que a biblioteca NLTK do Python tinha ferramentas e técnicas adequadas para a língua portuguesa.

Para concluir as questões, enunciadas no subcapítulo 1.4., foram todas respondidas e os objetivos propostos no ínicio desta dissertação foram atingidos.

6.1. Limitações do trabalho

Uma das limitações nesta investigação foi o facto de não existirem sistemas de extração de informação médica para língua portuguesa, pelo que houve a necessidade de criar um novo sistema para extrair especificamente a informação dos REM's do hospital português, objeto de estudo nesta pesquisa. Esta limitação dificultou a seleção dos termos clínicos mais relevantes e impediu a categorização dos termos em tipos de entidades clínicas, como por exemplo, em sintomas, medicação, procedimentos médicos, entre outros.

6.2. Trabalho Futuro

Na continuação deste trabalho foi pensado em desenvolver um dicionário médico em língua portuguesa, para extrair o conhecimento clínico de forma a melhorar a classificação de cada registo para hospitais portugueses. Após a melhoria da extração da informação médica e consequentemente a melhoria do classificador, pode ser desenhado um sistema não só para a extração de conhecimento estatístico, mas principalmente para ajudar o médico na sua tomada de decisão quanto ao diagnóstico correto para o paciente, criando sugestões de tipos de doença possíveis para determinados sintomas, resultados de exames clínicos, medicação, entre outros. Desta forma seria possível diminuir a taxa de erro do diagnóstico e naturalmente melhorar a qualidade do atendimento ao paciente.

Referências

[1] H. C. Koh and G. Tan, “Data mining applications in healthcare.,” J. Healthc. Inf. Manag., vol. 19, no. 2, pp. 64–72, 2005.

[2] J. C. Prather, D. F. Lobach, L. K. Goodwin, J. W. Hales, M. L. Hage, and W. E. Hammond, “Medical data mining: knowledge discovery in a clinical data warehouse.,” Proc. a Conf. Am. Med. Informatics Assoc. AMIA Fall Symp., pp. 101–5, 1997.

[3] F. W. Ms, S. Austin, and B. Mha, “Refereed papers The role of the electronic medical record (EMR) in care delivery development in developing countries : a systematic review,” pp. 139–146, 2008.

[4] J. Iavindrasana, G. Cohen, A. Depeursinge, H. Müller, R. Meyer, and A. Geissbuhler, “Clinical data mining: a review.,” Yearb. Med. Inform., pp. 121–133, 2009.

[5] M. J. Tierney, N. M. Pageler, M. Kahana, J. L. Pantaleoni, and C. A. Longhurst, “Medical Education in the Electronic Medical Record (EMR) Era : Benefits , Challenges , and Future Directions,” vol. 88, no. 6, pp. 748–752, 2013.

[6] I. Yoo et al., “Data mining in healthcare and biomedicine: A survey of the literature,” J. Med. Syst., vol. 36, no. 4, pp. 2431–2448, 2012.

[7] R. Hillestad et al., “Can Electronic Medical Record Systems Transform Health Care? Potential Health Benefits, Savings, And Costs,” pp. 1103–1117.

[8] I. H. Witten, “Text mining,” Pract. Handb. Internet Comput., pp. 14-1-14–22, 2004.

[9] A.-H. Tan, “Text Mining: The state of the art and the challenges,” Proc. PAKDD 1999 Work. Knowl. Disocovery from Adv. Databases, vol. 8, pp. 65–70, 1999.

[10] V. Korde, “Text Classification and Classifiers:A Survey,” Int. J. Artif. Intell. Appl., vol. 3, no. 2, pp. 85–99, 2012.

[11] F. Sebastiani and I. National, “Encyclopedia of Database Systems,” Encycl. Database Syst., no. March, pp. 0–5, 2009.

[12] F. De Comité, R. Gilleron, and M. Tommasi, “Learning Multi-label Alternating Decision Trees from Texts and Data,” in Machine Learning and Data Mining in Pattern Recognition, 2003, pp. 35–49.

[13] D. Capurro, M. Yetisgen, E. Eaton, R. Black, and P. Tarczy-Hornoch, “Availability of Structured and Unstructured Clinical Data for Comparative Effectiveness Research and Quality Improvement: A Multi-Site Assessment,” eGEMs (Generating Evid. Methods to Improv. patient outcomes), vol. 2, no. 1, p. 11, 2014.

[14] S. T. Rosenbloom, J. C. Denny, H. Xu, N. Lorenzi, W. W. Stead, and K. B. Johnson, “Data from clinical notes: A perspective on the tension between structure and flexible documentation,” J. Am. Med. Informatics Assoc., vol. 18, no. 2, pp. 181–186, 2011.

[15] Y. Wang et al., “Clinical information extraction applications: A literature review,” J. Biomed. Inform., vol. 77, no. November 2017, pp. 34–49, 2018.

[16] R. H. Miller and I. Sim, “Physicians ’ Use Of Electronic,” pp. 116–126, 2004.

[17] G. K. Savova et al., “Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications,” J. Am. Med. Informatics Assoc., vol. 17, no. 5, pp. 507–513, 2010.

[18] H. Analytics, “Electronic Medical Records vs . Electronic Health Records : Yes , There

Is a Difference By Dave Garets and Mike Davis Updated January 26 , 2006 HIMSS Analytics , LLC 230 E . Ohio St ., Suite 600 Chicago , IL 60611-3270 EMR vs . EHR : Definitions The marke," pp. 1–14, 2006.

[19] W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang, "Data processing and text mining technologies on electronic medical records: A review," J. Healthc. Eng., vol. 2018, 2018.

[20] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle, "Extracting information from textual documents in the electronic health record: a review of recent research.," Yearb. Med. Inform., pp. 128–144, 2008.

[21] P. Yadav, M. Steinbach, V. Kumar, and G. Simon, "Mining electronic health records (EHRs): A survey," ACM Comput. Surv., vol. 50, no. 6, pp. 1–40, 2018.

[22] L. Ohno-Machado, P. Nadkarni, and K. Johnson, "Natural language processing: Algorithms and tools to extract computable information from EHRs and from the biomedical literature," J. Am. Med. Informatics Assoc., vol. 20, no. 5, p. 805, 2013.

[23] R. Feldman, Y. Regev, E. Hurvitz, and M. Finkelstein-Landau, "Mining the biomedical literature using semantic analysis and natural language processing techniques," Drug Discov. Today BIOSILICO, vol. 1, no. 2, pp. 69–80, 2003.

[24] E. D. Liddy, "Natural Language Processing. In Encyclopedia of Library and Information Science," Marcel Decker, Inc., pp. 1–15, 2001.

[25] R. Rodriguez-Esteban, "Biomedical text mining and its applications," PLoS Comput. Biol., vol. 5, no. 12, pp. 1–5, 2009.

[26] C. Y. Wu et al., "Evaluation of Smoking Status Identification Using Electronic Health Records and Open-Text Information in a Large Mental Health Case Register," PLoS One, vol. 8, no. 9, pp. 1–8, 2013.

[27] K. Liu, K. J. Mitchell, W. W. Chapman, and R. S. Crowley, "Automating tissue bank annotation from pathology reports - comparison to a gold standard expert annotation set.," AMIA Annu. Symp. Proc., vol. 11, no. Figure 1, pp. 460–464, 2005.

[28] Q. T. Zeng, S. Goryachev, S. Weiss, M. Sordo, S. N. Murphy, and R. Lazarus, "Extracting principal diagnosis, co-morbidity and smoking status for asthma research: Evaluation of a natural language processing system," BMC Med. Inform. Decis. Mak., vol. 6, pp. 1–9, 2006.

[29] K. P. Liao et al., "Electronic medical records for discovery research in rheumatoid arthritis," Arthritis Care Res., vol. 62, no. 8, pp. 1120–1127, 2010.

[30] B. E. Himes, Y. Dai, I. S. Kohane, S. T. Weiss, and M. F. Ramoni, "Prediction of Chronic Obstructive Pulmonary Disease (COPD) in Asthma Patients Using Electronic Medical Records," J. Am. Med. Informatics Assoc., vol. 16, no. 3, pp. 371–379, 2009.

[31] G. K. Savova et al., "DeepPhe: A natural language processing system for extracting cancer phenotypes from clinical records," Cancer Res., vol. 77, no. 21, pp. e115–e118, 2017.

[32] G. K. Savova et al., "Discovering peripheral arterial disease cases from radiology notes using natural language processing," AMIA Annu. Symp. Proc., vol. 2010, pp. 722–726, 2010.

[33] A. N. Ananthakrishnan et al., "Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: A novel informatics approach," Inflamm. Bowel Dis., vol. 19, no. 7, pp. 1411–1420, 2013.

[34] A. Khalifa and S. Meystre, "Adapting existing natural language processing resources for

cardiovascular risk factors identification in clinical notes," J. Biomed. Inform., vol. 58, pp. S128–S132, 2015.

[35] J. K. Hou et al., "Automated identification of surveillance colonoscopy in inflammatory bowel disease using natural language processing," Dig. Dis. Sci., vol. 58, no. 4, pp. 936–941, 2013.

[36] H. Salmasian, D. E. Freedberg, and C. Friedman, "Deriving comorbidities from medical records using natural language processing," J. Am. Med. Informatics Assoc., vol. 20, no. E2, pp. 239–242, 2013.

[37] H. S. Chase, L. R. Mitrani, G. G. Lu, and D. J. Fulgieri, "Early recognition of multiple sclerosis using natural language processing of the electronic health record," BMC Med. Inform. Decis. Mak., vol. 17, no. 1, p. 24, 2017.

[38] N. L. Jain and C. Friedman, "Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports.," Proc. a Conf. Am. Med. Informatics Assoc. AMIA Fall Symp., pp. 829–33, 1997.

[39] and E. A. M. André Coutinho Castillaa, Sérgio Shiguemi Furuiea, "Multilingual Information Retrieval in Thoracic Radiology: Feasibility Study," Stud Heal. Technol Inform. 2007 ; 129(0 1) 387–391.

[40] G. Schadow and C. J. McDonald, "Extracting structured information from free text pathology reports.," AMIA Annu. Symp. Proc., pp. 584–588, 2003.

[41] M. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text classification using machine learning techniques," WSEAS Trans. Comput., vol. 4, no. 8, pp. 966–974, 2005.

[42] M. Aly, "Survey on Multiclass Classification Methods Extensible algorithms," no. November, pp. 1–9, 2005.

[43] M. K. Dalal and M. A. Zaveri, "Automatic Text Classification: A Technical Review," Int. J. Comput. Appl., vol. 28, no. 2, pp. 37–40, 2011.

[44] A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining," Brief. Bioinform., vol. 6, no. 1, pp. 57–71, 2005.

[45] A. M. Cohen, "An effective general purpose approach for automated biomedical document classification.," AMIA Annu. Symp. Proc., pp. 161–165, 2006.

[46] H. Tou, L. Yao, Z. Wei, X. Zhuang, and B. Zhang, "Automatic infection detection based on electronic medical records," BMC Bioinformatics, vol. 19, no. Suppl 5, 2018.

[47] S. Pakhomov, N. Shah, P. Hanson, S. Balasubramaniam, and S. A. Smith, "Automatic Quality of Life Prediction Using Electronic Medical Records," pp. 545–549, 2008.

[48] S. V. S. Pakhomov, P. L. Hanson, S. S. Bjornsen, and S. A. Smith, "Automatic Classification of Foot Examination Findings Using Clinical Notes and Machine Learning," J. Am. Med. Informatics Assoc., vol. 15, no. 2, pp. 198–202, 2008.

[49] J. Munroe, "Automatic Methods to Extract New York Heart Association Classification from Clinical Notes," pp. 0–3, 2017.

[50] A. Tharwat, "Classification assessment methods," Appl. Comput. Informatics, 2018.

[51] M. Grandini, E. Bagli, and G. Visani, "Metrics for Multi-Class Classification: an Overview," pp. 1–17, 2020.

[52] H. M and S. M.N, "A Review on Evaluation Metrics for Data Classification Evaluations," Int. J. Data Min. Knowl. Manag. Process, vol. 5, no. 2, pp. 01–11, 2015.

[53] J. M. Gálvez et al., “Multiclass classification for skin cancer profiling based on the integration of heterogeneous gene expression series,” PLoS One, vol. 13, no. 5, pp. 1–26, 2018.

[54] D. M. Farid, L. Zhang, C. M. Rahman, M. A. Hossain, and R. Strachan, “Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks,” Expert Syst. Appl., vol. 41, no. 4 PART 2, pp. 1937–1946, 2014.

[55] C. Mercan, S. Aksoy, E. Mercan, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, “Multi-Instance Multi-Label Learning for Multi-Class Classification of Whole Slide Breast Histopathology Images,” IEEE Trans. Med. Imaging, vol. 37, no. 1, pp. 316–325, 2018.

[56] K. M. Schneider, “On word frequency information and negative evidence in naive bayes text classification,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 3230, pp. 474–485, 2004.

[57] E. Frank and R. R. Bouckaert, “Naive bayes for text classification with unbalanced classes,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 4213 LNAI, pp. 503–510, 2006.

[58] S. Bansal, C. Gupta, and A. Arora, “User tweets based genre prediction and movie recommendation using LSI and SVD,” 2016 9th Int. Conf. Contemp. Comput. IC3 2016, 2017.

Anexo A

In [36]:
```
#Importação de Bibliotecas

import csv
import nltk
from nltk.corpus import stopwords
from pprint import pprint
import spacy
from scipy import sparse
from scipy.sparse import csr_matrix
from scipy.sparse import csc_matrix
from scipy.sparse import coo_matrix
import sklearn
import re
import pandas as pd
from sklearn.naive_bayes import GaussianNB
from sklearn import svm
from sklearn.metrics import accuracy_score
from nltk.classify import NaiveBayesClassifier
from sklearn.feature_extraction.text import TfidfVectorizer
from scipy.sparse import coo_matrix, hstack
from sklearn.metrics import recall_score
from sklearn.metrics import precision_score
from sklearn.metrics import f1_score
from nltk.stem import PorterStemmer
from nltk.corpus import stopwords
```

I. Pré-Processamento dos dados

In [37]:
```
#Importação do ficheiro excel que contem os dados médicos

excel = pd.read_excel("HDI-Episodios com diagnosticos e diario clinico.xlsx")
```

In [38]:
```
#Remoção dos missing values

excel1 = excel.dropna(how='any',axis=0)
```

In [39]:
```
#Seleção das únicas colunas que vamos utilizar para o estudo

df1 = excel1[['DES_DIAGNOSTICO', 'DIARIO']]
```

In [40]:
```
#Selação dos 4 diagnósticos com mais registos e por isso melhor capacidade preditiva

dfinal = df1.loc[(df1['DES_DIAGNOSTICO'] == 'DOENCAS DO SANGUE E ORGAOS HEMATOP.') | (df1['DES_DIAGNOSTICO'] == 'DOENCAS DO SISTEMA OSTEOART. E TECIDOS CONJUNT.') | (df1['DES_DIAGNOSTICO'] == 'DOENCAS INFECCIOSAS E PARASITARIAS') | (df1['DES_DIAGNOSTICO'] == 'TUMORES (NEOPLASMAS)')]


#Análise dos resgistos após esta seleção

print(dfinal.info())
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9487 entries, 580 to 11169
Data columns (total 2 columns):
DES_DIAGNOSTICO    9487 non-null object
DIARIO             9487 non-null object
dtypes: object(2)
memory usage: 222.4+ KB
None
```

In [41]:
```
#Aplicar um Random à posição dos registos

s = dfinal.sample(frac=1)
```

In [42]:
```
#Passar os dados que estão num dataframe para o formato de lista

DES_DIAGNOSTICO = s['DES_DIAGNOSTICO'].values.tolist()
```

In [43]:
```
#Passar os dados que estão num dataframe para o formato de lista

DIARIO = s['DIARIO'].tolist()
```

In [44]:
```
#Definição de uma função para aplicar o formato Lista a cada registo do diário

def string():
    DIARIO1 = []

    for e in DIARIO:
        aaa = str(e)
        DIARIO1.append(aaa)

    return DIARIO
DIARIO1 = string()
```

II. Processamento NLP

(Minúsculas, Tokanização, Remoção de Stopwords e Pontuação, Stemming, POS-Tag, TF-IDF)

In [45]:
```
#Definição da função para colocar todas as letras em minúsculo

def minusculas(lista):

    lower_list = []

    for text in lista:

        lower_list.append(str(text).lower())

    return lower_list

DIARIO_en_lowercase = minusculas(DIARIO1)
```

In [46]:
```
#Definição da função de tokenization para colocar as frases num conjunto de to
kens

def tokanizacao(lista):
    documentos_tokanizados = []

    for text in lista:

        expressao_reg = re.sub(r"([\-\.\?\!\;\:\`\'\,])", r" \1 ", text)
        #expressao_reg2 = re.sub(r"(('s))", r" . ", expressao_reg)
        #expressao_reg3 = re.sub(r"((`))", r".", expressao_reg2)
        #expressao_reg4 = re.sub('[^a-z]', ' ', expressao_reg2)
        new_text = nltk.word_tokenize(expressao_reg)
        documentos_tokanizados.append(new_text)

    return documentos_tokanizados

DIARIO_en_tokens = tokanizacao(DIARIO_en_lowercase)
```

In [47]:
```
#Definição da função do Stemming para reduzir todas as palavras as seu radical
e remoção das stopwords

ps = PorterStemmer()
docTotal = []
stopw = stopwords.words('portuguese')

def stemming(lista):

    for ste in lista:
        doc = []
        for w in ste:
            if w.lower() not in stopw:
                stemmer = ps.stem(w)
                doc.append(stemmer)
        docTotal.append((doc))

    return docTotal

DIARIO_ste = stemming(DIARIO_en_tokens)
```

In [48]:
```
#Definição da função POS-TAG para atribuir a cada palavra a sua classe gramati
cal
#Foram adicionados numas palavras à lista de stopwords que não demonstram nenh
um
#interesse para a análise e foram removidas essas palavras
#Foram também selecionados apenas as palavras cuja classe gramatical eram nome
s comuns e próprios no singular e no plural

def pos_tag(lista):

    POSTAG_INGLES = [nltk.pos_tag(w) for w in lista]
    stopw_english = stopwords.words('portuguese')
    stopw_english.append('u')
    stopw_english.append('k')
    stopw_english.append('l')
    stopw_english.append('q')
    stopw_english.append('w')
    stopw_english.append('e')
    stopw_english.append('r')
    stopw_english.append('s')
    stopw_english.append('t')
    stopw_english.append('u')
    stopw_english.append('i')
    stopw_english.append('o')
    stopw_english.append('p')
    stopw_english.append('a')
    stopw_english.append('d')
    stopw_english.append('f')
    stopw_english.append('g')
    stopw_english.append('h')
    stopw_english.append('i')
    stopw_english.append('j')
    stopw_english.append('l')
    stopw_english.append('m')
    stopw_english.append('z')
    stopw_english.append('x')
    stopw_english.append('c')
    stopw_english.append('v')
    stopw_english.append('b')
    stopw_english.append('n')
    stopw_english.append('pnrp')
    stopw_english.append('inf')
    stopw_english.append("es")
    stopw_english.append("ml")

    docs_taggados = []
    for noticia in POSTAG_INGLES:
        aux3 = []
        for word,pos in noticia:
            if pos == 'NNS' or pos=='NN' or pos == 'NNP' or pos == 'NNPS' and
word not in stopw_english:
                aux3.append(word)
        docs_taggados.append(aux3)

    docs_taggados
```

```
    return docs_taggados

DIARIO_en_postag = pos_tag(DIARIO_ste)
```

In [49]:
```
#Seleção das features mais importantes utilizando o TF-IDF
#Foram apenas selecionados 1000 features

final = []
for palavras in DIARIO_en_postag:
    aux = ''
    for palavra in palavras:
        inicio = ' '
        medio = inicio + palavra
        aux = aux + medio
    final.append(aux)

vectorizer = TfidfVectorizer(max_features=1000)
join_idf = vectorizer.fit_transform(final)
#max_features=5000
```

In [50]:
```
#Matriz

join_idf
```

Out[50]:
```
<9487x1000 sparse matrix of type '<class 'numpy.float64'>'
        with 122514 stored elements in Compressed Sparse Row format>
```

III. Classificação

In [67]:
```
#Divisão do dataset em treino e teste

train = int(len(s) * 80 / 100)

train_x, test_x = (join_idf[:train], join_idf[train:])

train_y, test_y = (DES_DIAGNOSTICO[:train], DES_DIAGNOSTICO[train:])
```

In [68]:
```
#Aplicação do algoritmo Naive Bayes Bernoulli

from sklearn.naive_bayes import BernoulliNB
clf = BernoulliNB()
final = clf.fit(train_x, train_y)
Y_pred = final.predict(test_x)
print("Accuracy: " + str(accuracy_score(Y_pred, test_y)))
print("Recall: " + str(recall_score(test_y, Y_pred, average='micro')))
print("Precision: " + str(precision_score(test_y, Y_pred, average='micro')))
print("F1: " + str(f1_score(test_y, Y_pred, average='micro')))
```

```
Accuracy: 0.8877766069546892
Recall: 0.8877766069546892
Precision: 0.8877766069546892
F1: 0.8877766069546892
```

In [71]:
```
#Aplicação do algoritmo Naive Bayes Gaussian

from sklearn.naive_bayes import GaussianNB
clf = GaussianNB()
final = clf.fit((train_x.toarray()), train_y)
Y_pred = final.predict((test_x.toarray()))
print("Accuracy: " + str(accuracy_score(Y_pred, test_y)))
print("Recall: " + str(recall_score(test_y, Y_pred, average='micro')))
print("Precision: " + str(precision_score(test_y, Y_pred, average='micro')))
print("F1: " + str(f1_score(test_y, Y_pred, average='micro')))
```

```
Accuracy: 0.8598524762908325
Recall: 0.8598524762908325
Precision: 0.8598524762908325
F1: 0.8598524762908325
```

In [72]:
```
#Aplicação do algoritmo Naive Bayes Multinomial - ESCOLHIDO - APRESENTOU MELHO
RES RESULTADOS

from sklearn.naive_bayes import MultinomialNB
clf = MultinomialNB()
final = clf.fit(train_x, train_y)
Y_pred = final.predict(test_x)
print(Y_pred[1000])
print(test_y[1000])
print("Accuracy: " + str(accuracy_score(Y_pred, test_y)))
print("Recall: " + str(recall_score(test_y, Y_pred, average='weighted')))
print("Precision: " + str(precision_score(test_y, Y_pred, average='weighted'
)))
print("F1: " + str(f1_score(test_y, Y_pred, average='weighted')))
```

```
TUMORES (NEOPLASMAS)
DOENCAS DO SANGUE E ORGAOS HEMATOP.
Accuracy: 0.8998946259220232
Recall: 0.8998946259220232
Precision: 0.8989420548758067
F1: 0.8993069331437661
```

In [76]:
```
#Aplicação das métricas ao último algoritmo aplicado que é o Naive Bayes Multi
nomial

from sklearn import metrics
print(metrics.classification_report(test_y, Y_pred, digits=3))
```

```
                                                    precision    recall  f1-scor
e   support

            DOENCAS DO SANGUE E ORGAOS HEMATOP.      0.809     0.779     0.79
4       403
DOENCAS DO SISTEMA OSTEOART. E TECIDOS CONJUNT.      0.963     0.965     0.96
4       399
             DOENCAS INFECCIOSAS E PARASITARIAS      0.935     0.970     0.95
2       298
                           TUMORES (NEOPLASMAS)      0.899     0.902     0.90
1       798

                                       accuracy                          0.90
0      1898
                                      macro avg      0.901     0.904     0.90
3      1898
                                   weighted avg      0.899     0.900     0.89
9      1898
```