INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA INFORMÁTICA




Elaboração Automática de Relatórios Médicos

ALESSANDRO FELIZ DOS SANTOS
outubro de 2017

POLITÉCNICO
DO PORTO

Instituto Superior de Engenharia do Porto

Elaboração Automática de Relatórios Médicos

Alessandro Feliz dos Santos

Dissertação para obtenção do Grau de Mestre em Engenharia Informática

Orientadora: Goreti Marreiros

Supervisor: Ricardo Freitas

Porto, outubro 2017

«Dedico este trabalho a todos os que me ajudaram e apoiaram durante a minha vida.»

Resumo

A presente dissertação foi realizada num contexto empresarial na empresa FUJIFILM. Foi identificado um problema real e comum a muitos hospitais, o gasto excessivo de tempo e dinheiro no processo de elaboração de relatórios clínicos. Face este problema, apresenta-se uma forma alternativa de elaboração de relatórios que, passa pela previsão de texto mediante as palavras redigidas pelo utilizador. A solução proposta permite em primeiro lugar, diminuir o tempo da redação de relatórios, disponibilizando ao utilizador sugestões para o resto das suas frases. Em segundo lugar, permite reduzir os custos associados à contratação de técnicos de transcrição e compra de *hardware*/*software* necessários para a transcrição automática. Finalmente, em terceiro lugar, permitirá uma redução da ocorrência de erros ortográficos através da diminuição da quantidade de texto redigido. Durante o estudo do estado da arte foram identificadas três abordagens possíveis para a implementação do conceito idealizado, sendo elas: abordagens baseadas em regras; abordagens baseadas em estatística e probabilidade; e abordagens hibridas. De todos os métodos incluídos nas abordagens identificadas, destacam-se os três mais utilizados na previsão de texto, sendo eles: Redes Neuronais; *NGrams*; e *Suffix Trees*. Apôs analisadas as vantagens e desvantagens de cada abordagem e respetivos métodos, optou-se pela implementação de um sistema baseado no conceito de *NGrams*, principalmente pelos resultados positivos obtidos por outros autores e facilidade de implementação. É apresentado depois o desenho e implementação da solução proposta, onde se aborda o problema da *performance*, respetiva resolução e a configuração do sistema. Finamente, é apresentada a avaliação da solução desenvolvida, onde definimos para avaliação as seguintes métricas: tempo de redação de um relatório; percentagem de *keystroke savings*; e tempo médio de uma previsão. A partir destas métricas foi depois possível a formulação das hipóteses a testar, tendo-se obtido *keystroke savings* entre os 69 e os 90%, e tempos médios de previsão entre os 42 e 2 226 milisegundos, dependendo da configuração do sistema.

Palavras-chave: Previsão de texto, *language models*, *NGrams*

Abstract

This dissertation was carried out in a business context in the company FUJIFILM. A real problem common to many hospitals was identified, consisting in the excessive waste of time and money in the process of writing clinical reports. Given this problem, an alternative approach for the elaboration of medical reports is presented, involving the prediction of text using the previously written words. The proposed solution allows first of all, to reduce the time spent in the process of writing reports, providing to the user suggestions for the rest of his sentences. Secondly, it will decrease the costs associated with the hiring of transcription technicians and prevents the need of buying software/hardware for the automatic transcription. Thirdly, it will decrease the occurrence of spelling errors by reducing the quantity of text written by the user. During the study of the state of the art three possible approaches for the implementation of the idealized concept were identified: approaches based on rules; approaches based in statistics and probabilities; and hybrid approaches. Of all the methods included in the previous identified approaches, three are worth mentioning: Neural Networks; NGrams; and Suffix Trees. After analyzing the advantages and disadvantages of each approach and method previously presented, we concluded that the best method to use would be the NGrams, primarily due to the positive results obtained by other authors and easy implementation. After that we present the design and implementation of the proposed system, were we also talk about the performance issues found and system configurations. Finally, we present to the reader the evaluation of the developed solution, where we defined the following metrics for the evaluation: time taken to write a report; keystroke savings; and prediction time. Overall the developed system achieved keystroke savings values ranging from 69 to 90% and prediction times between 42 and 2 226 milliseconds, depending on the configuration used.

Keywords: Text prediction, language models, *NGrams*

Agradecimentos

Quero agradecer a todos aqueles que me ajudaram e apoiaram, não só durante a realização desta dissertação, como também durante a realização do mestrado.

Em primeiro lugar, quero agradecer à minha família, principalmente aos meus pais pela motivação dada ao longo destes anos todos e por me terem tornado na pessoa que sou hoje.

Um obrigado especial aos meus supervisores Élio Santos e Ricardo Freitas pela oportunidade oferecida de desenvolver este projeto na FUJIFILM.

Um obrigado também especial a minha orientadora, Doutora Goreti Marreiros, por ter aceite ajudar-me nesta fase importante da minha vida.

Gostava também de agradecer ao ISEP e a todos os docentes do DEI pelos conhecimentos transmitidos ao longo dos anos.

E por último, devo também uma palavra de apreço à minha namorada, Inês, por ter escutado todos os meus desabafos.

A todos, um sincero obrigado.

Índice

1 Introdução 1

1.1 Âmbito 1

1.2 FUJIFILM 1

1.3 Contexto 3

1.3.1 Técnicas de elaboração de relatórios 4

1.3.2 Intervenientes 4

1.3.3 Normas de interoperabilidade 5

1.4 Problema 6

1.5 Objetivos 7

1.6 Estrutura do relatório 7

2 Análise de valor 9

2.1 Produto alvo 9

2.2 Análise de valor 9

2.2.1 Valor 11

2.2.2 Modelo Canvas 11

3 Enquadramento teórico 15

3.1 Áreas abrangidas 15

3.2 Conceitos 16

3.3 Estado da arte 19

3.3.1 Origem 19

3.3.2 Atualidade 20

3.4 Aplicações de previsão de texto 22

4 *Design* 25

4.1 Tecnologias 25

4.2 Análise de possíveis soluções 25

4.3 *Design* 29

4.3.1 Requisitos 29

4.3.2 Arquitetura 31

5 Desenvolvimento 37

5.1 Desenvolvimento 37

5.2 *Performance* 41

5.2.1 Memória 41

5.2.2 Processamento 41

5.3 Configuração do sistema 42

6 Avaliação 43
6.1 Métricas 43
6.2 Hipóteses 44
6.3 Metodologia de avaliação 45
6.4 Resultados 46
6.5 Análise dos resultados 51
7 Conclusão 55
7.1 Resumo 55
7.2 Objetivos realizados 56
7.3 Limitações e trabalho futuro 56
7.4 Apreciação final 57
Referências 59

Lista de Figuras

Figura 1 – Segmentos de mercado onde a FUJIFILM opera 2
Figura 2 – *Workflow* de trabalho no serviço de radiologia. 3
Figura 3 – *New Concept Development*. 10
Figura 4 – Modelo canvas. 13
Figura 5 – Processo completo de treino (sem detalhes). 16
Figura 6 – Processo completo de aprendizagem (sem detalhes). 16
Figura 7 – RN para previsão de texto. 19
Figura 8 – Teclado preditivo da VocabPC. 23
Figura 9 – Árvore de decisão. 26
Figura 10 – Diagrama de *Use Cases*. 30
Figura 11 – *Flow* do CWM *Reporting* com o sistema de previsão. 32
Figura 12 – Etapas do pré processamento. 34
Figura 13 – Arquitetura do sistema. 35
Figura 14 – Diagrama de atividades relativo ao processo de previsão. 40

Lista de Tabelas

Tabela 1 – Resumo das abordagens baseadas em regras. 20
Tabela 2 – Resumo das abordagens baseadas em estatística e probabilidades. 21
Tabela 3 – Comparação de diversos fatores entre as várias abordagens possíveis. 26
Tabela 4 – Escala fundamental de Saaty. 27
Tabela 5 – Comparação de critérios. 27
Tabela 6 – Comparação entre critérios normalizada e prioridade relativa. 27
Tabela 7 – Quadro resumo das configurações testadas. 46
Tabela 8 – Resultados obtidos com a configuração A1. 47
Tabela 9 – Resultados obtidos com a configuração A2. 47
Tabela 10 – Resultados obtidos com a configuração B1. 48
Tabela 11 – Resultados obtidos com a configuração B2. 48
Tabela 12 – Resultados obtidos com a configuração C1. 49
Tabela 13 – Resultados obtidos com a configuração C2. 49
Tabela 14 – Resultados obtidos com a configuração D1. 50
Tabela 15 – Resultados obtidos com a configuração D2. 50
Tabela 16 – Quadro resumo dos resultados obtidos. 51
Tabela 17 – Grau de conclusão dos objetivos propostos. 56
Tabela 18 – Valores de comparação entre abordagens no critério complexidade. 64
Tabela 19 – Valores de comparação do critério complexidade normalizados e prioridades. 64
Tabela 20 – Valores de comparação entre abordagens no critério AA. 64
Tabela 21 – Valores de comparação do critério AA normalizados e prioridades. 64
Tabela 22 – Valores de comparação entre abordagens no critério requer dados. 65
Tabela 23 – Valores de comparação do critério requer dados normalizados e prioridades. 65
Tabela 24 – Valores de comparação entre abordagens no critério requer dados. 65
Tabela 25 – Valores de comparação do critério resultados normalizados e prioridades. 65

Lista de Gráficos

Gráfico 1 – KS dos títulos, corpos e documentos completos para todas as configurações. 52
Gráfico 2 – Tempos médios (em milisegundos) das previsões para todas as configurações. ... 53

Acrónimos e Símbolos

Lista de Acrónimos

AA	Aprendizagem Automática
AHP	*Analytic Hierarchy Process*
BIS	*Breast Information System*
CEOM	Conselho Europeu das Ordens dos Médicos
CWM	*Clinical Workflow Manager*
DICOM	*Digital Imaging and Communications in Medicine*
DLL	*Dynamic Link Library*
FFE	*Fuzzy Front End*
GC	*Garbage Collector*
GIS	*Gastroenterology Information System*
HL7	*Health Level 7*
IA	Inteligência Artificial
IC	Índice de Consistência
IR	Índice Aleatório
IBM	*International Business Machines*
ISEP	Instituto Superior de Engenharia do Porto
KS	*Keystroke Savings*
KSPC	*Keystrokes per Character*
LM	*Language Modeling*
ML	*Machine Learning*
MEI	Mestrado de Engenharia Informática
NCD	*New Concept Development model*
NPD	*New Product Development*

PACS	*Picture Archiving and Communication System*
PLN	Processamento de Linguagem Natural
RC	Razão de Consistência
RF	Requisitos Funcionais
RIS	*Radiology Information System*
RN	Redes Neuronais
RNF	Requisitos Não Funcionais
SIH	Sistemas de Informação Hospitalares
SMS	*Short Message Service*
SNOMED CT	*Systematized Nomenclature of Medicine Clinical Terms*
TFS	*Team Foundation Server*
VS15	*Visual Studio 2015*

Lista de Símbolos

P	Probabilidade
F	Frequência
N	Número de *tokens*
C	Tamanho do contexto
T	*Token*
α	Vetor prioridade
γ_{max}	Valor próprio

1 Introdução

Neste capítulo é realizada uma breve introdução a presente dissertação. Inicialmente é apresentado o âmbito e a empresa onde a mesma foi desenvolvida. De seguida, é exposto ao leitor todo o ambiente em redor do problema que se procura resolver, sendo inicialmente apresentado o seu contexto, os processos e intervenientes associados, o problema em si, e os objetivos a atingir. No final é ainda apresentada a estrutura deste documento.

1.1 Âmbito

A presente dissertação foi desenvolvida no âmbito do Mestrado de Engenharia Informática (MEI), no Instituto Superior de Engenharia do Porto (ISEP), sendo o seu objetivo a atribuição do grau de mestre ao autor na área de Engenharia Informática. O trabalho realizado nesta dissertação foi desenvolvido num contexto empresarial consistindo, o mesmo, na resolução de um problema real proposto pela FUJIFILM. No final, dependendo dos resultados obtidos, existirá a possibilidade de a solução desenvolvida ser incorporada em alguns dos produtos já comercializados pela FUJIFILM.

1.2 FUJIFILM

Como referido anteriormente, esta dissertação foi realizada com base num projeto proposto pela FUJIFILM, tratando-se de uma empresa multinacional japonesa, fundada em 1934 e que começou por desenvolver produtos fotográficos, nomeadamente, películas fotográficas (FUJIFILM, 2017).

Com o passar dos anos, a empresa começou a expandir-se para outras áreas de negócio, encontrando-se atualmente presente nos mais variados mercados relacionados com sistemas de informação e imagem. Na figura 1 é possível ver a distribuição da FUJIFILM nos vários segmentos de mercado em que opera. A informação apresentada reporta ao ano fiscal de 2016.

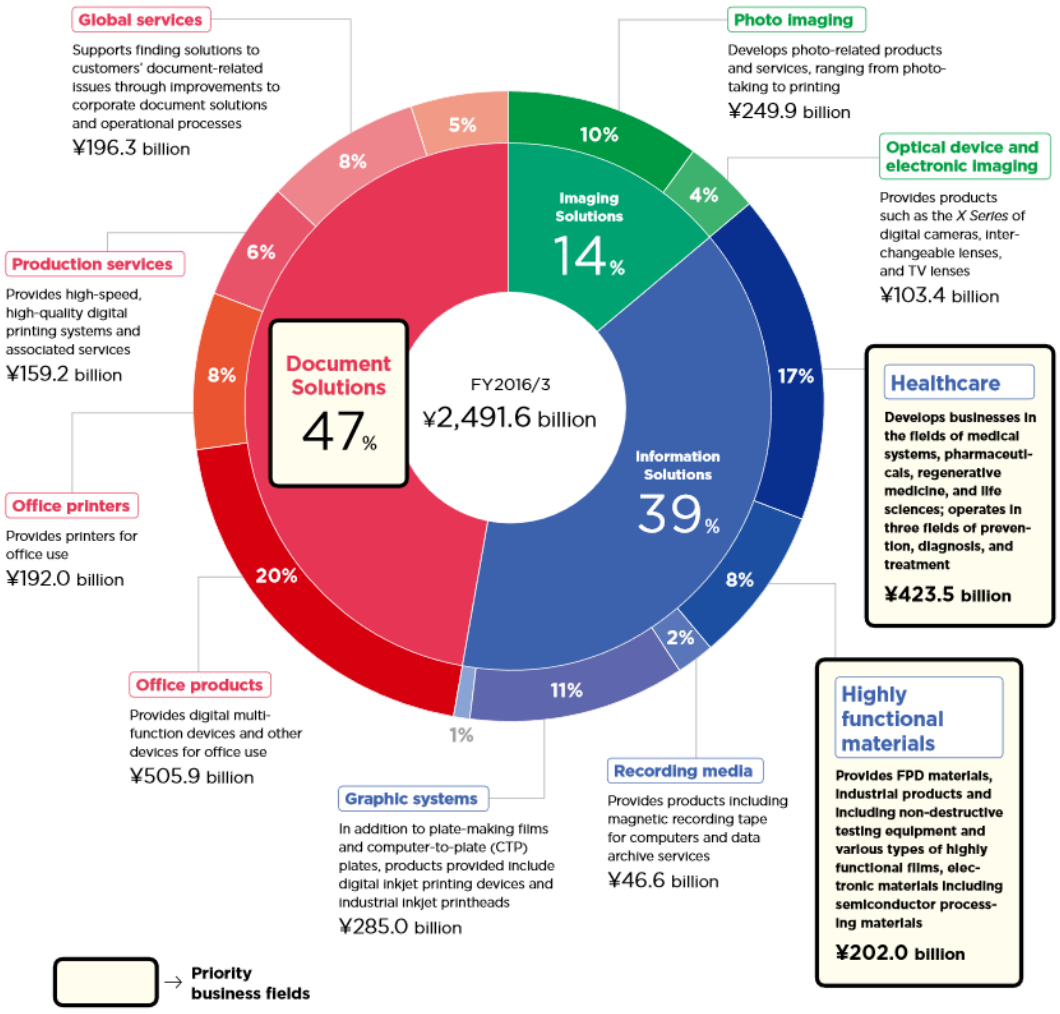

Figura 1 – Segmentos de mercado onde a FUJIFILM opera (FUJIFILM, 2016).

Na figura apresentada é possível identificar três mercados principais: sistemas para documentação, ocupando 47% do universo de negócios da FUJIFILM; sistemas de informação, 39%; e sistemas de imagens, 14%. É de realçar que o mercado onde a empresa se iniciou, imagem, é hoje um dos que tem menor representação. Na mesma figura, é ainda possível ver o volume total de vendas, que é de aproximadamente 20,7 mil milhões de euros.

Relativamente à sucursal portuguesa, o departamento de desenvolvimento e investigação desenvolve primariamente soluções para a área da saúde, correspondendo a 17% do universo de negócios da FUJIFILM, com um volume total de vendas de 3,5 mil milhões de euros. O produto que se encontra atualmente a ser desenvolvido e onde a solução a desenvolver poderá ser integrada será apresentado com detalhe no capítulo 2.1.

1.3 Contexto

O aumento da população mundial (United Nations, 2015), assim como a rápida evolução tecnológica a que estamos suscetíveis nos dias de hoje, obrigam as instituições de saúde a um maior rigor no armazenamento de informações clínicas. Assim, cada vez mais assistimos a uma transição dos registos clínicos em papel para registos digitais (Martins, 2011). Tal obriga a um crescimento dos Sistemas de Informação Hospitalares (SIH), o que implica também que os mesmos fiquem cada vez mais complexos, e que cubram as várias necessidades dos hospitais, que vão desde a "*front-end*", que incluí a marcação/realização de exames, gestão de horários, recursos humanos, e muito mais, mas também o "*back-end*", que suporta todas estas tarefas.

Atualmente, qualquer evento clínico num contexto hospitalar, desde uma simples consulta de rotina a uma complexa intervenção cirúrgica, envolve um conjunto de passos que devem ser bem geridos de forma a não criar o caos nos serviços de saúde. Por exemplo, para a realização de uma radiografia é necessário, em primeiro lugar, a marcação do exame tendo em conta a disponibilidade do equipamento e do técnico de radiologia. De seguida, deve ser registada a chegada do paciente ao serviço de radiologia, devendo depois ser realizado o exame pelo técnico e equipamento designados na marcação. A próxima fase consiste no envio das imagens capturadas para um médico especialista, estas imagem não devem ser perdidas ou misturadas com as de outros pacientes. Após a análise das mesmas, o médico deve proceder à elaboração do relatório clínico. O processo de redação do relatório pode ser realizado de diversas formas (abordadas com maior detalhe no subcapítulo 1.3.1), mas para o presente exemplo podemos assumir que o médico procede à narração da sua análise, sendo a mesma gravada, e que posteriormente essa gravação é transcrita para texto. Finalmente, o relatório elaborado é entregue a quem efetuou o pedido do exame, como por exemplo, o médico de família, ou o médico de urgência.

Figura 2 – *Workflow* de trabalho no serviço de radiologia (FUJIFILM, 2016).

No *workflow* acima descrito identificam-se inúmeros pontos onde um simples erro pode trazer consequências graves para um hospital, desde o aumento do tempo de espera devido a marcações incorretas, aos diagnósticos errados devido a trocas de informações. Com isto, é possível compreender a importância dos sistemas de informação na gestão de um hospital e respetivo controlo de todas estas fases.

No âmbito da presente dissertação, iremo-nos focar nas fases relacionadas com a elaboração dos relatórios clínicos. Estes documentos podem ser utilizados em diversas situações, como: análises de risco para seguros de saúde; realização de autópsias; diagnósticos; entre outros. Tratam-se por isso de documentos de extrema importância, que devem ser redigidos com o maior rigor e exatidão possíveis por um especialista da área da saúde. Por exemplo, a realização

de um qualquer exame implica a posterior descrição do quadro clínico do paciente. Se algo de anormal for encontrado, um diagnóstico é realizado de forma a detetar a origem do problema e, se necessário, iniciar o respetivo tratamento. O diagnóstico consiste na análise da descrição do quadro clínico do paciente. Se essa descrição não for correta ou completa, tal pode dar origem a um diagnóstico errado, que poderá ter consequências graves, daí a importância da correta elaboração de um relatório clínico.

1.3.1 Técnicas de elaboração de relatórios

Nos dias de hoje existem três formas principais de elaborar relatórios clínicos, que podem ser utilizadas separadamente ou em conjunto, sendo elas:

- A partir de *templates* estáticos;
- Através da gravação áudio e posterior transcrição manual para texto;
- Através da gravação áudio e posterior transcrição automática para texto.

A utilização de *templates* estáticos baseia-se, como o próprio nome indica, na utilização de *templates* previamente criados que são depois editados e preenchidos pelos médicos. Estes documentos podem ser impressos, sendo depois preenchidos manualmente, ou então, editados diretamente no computador. Relativamente à abordagem baseada na transcrição manual, a mesma implica que o médico descreva um exame para um microfone. A gravação é depois armazenada sendo posteriormente convertida manualmente para texto por um técnico de transcrição. Quanto à transcrição automática, a mesma é muito similar a manual, sendo que a única diferença reside no transcritor, que neste caso é um sistema e não uma pessoa.

Todas as técnicas apresentadas são vastamente utilizadas, sendo que a escolha da abordagem adotada pelo hospital depende de diversos fatores, tais como: a quantidade de relatórios diários; o(s) serviço(s) hospitalar(es) em causa; os recursos financeiros do hospital; entre outros. Por exemplo, um hospital com poucos recursos tenderá a adotar uma abordagem baseada em *templates* estáticos, enquanto que um hospital numa situação oposta, utilizará uma das técnicas baseadas na transcrição.

Mesmo com a sua ampla utilização, qualquer uma destas técnicas apresenta problemas limitativos que são importantes de referir, sendo os mesmos abordados no subcapítulo 1.4.

1.3.2 Intervenientes

Face ao que até agora foi exposto, é possível identificar três intervenientes nos vários processos de elaboração de um relatório clínico: médicos; pacientes; e técnicos de transcrição. Os médicos e pacientes são os nossos atores principais e encontram-se em qualquer uma das técnicas anteriormente apresentadas. Segundo a *Carta Europeia de Ética Médica*, um médico é uma pessoa que usa os conhecimentos adquiridos durante a sua formação para melhorar, ou pelo

menos manter, a saúde dos seus pacientes. Com isto, a sua função principal passa pela prestação de cuidados de saúde (Conselho Europeu das Ordens dos Médicos (CEOM), 2011). No contexto apresentado, o médico será a pessoa responsável pela elaboração dos relatórios médicos. O paciente, como já referido, trata-se de uma pessoa que está a ser alvo de cuidados médicos, tratando-se do indivíduo alvo dos relatórios clínicos. Relativamente ao técnico de transcrição, tal como o nome indica, a sua função consiste na transcrição de áudio para texto. Este interveniente entra em cena numa das técnicas apresentadas, transcrevendo manualmente gravações áudio dos médicos para documentos de texto.

1.3.3 Normas de interoperabilidade

Devido à grande diversidade de sistemas de informação utilizados em diferentes hospitais, ou mesmo dentro do mesmo hospital, e ainda à grande quantidade de equipamentos médicos disponíveis, muitas das vezes de fabricantes diferentes, onde cada um seguia a sua própria norma para a gestão da informação, surgiu a necessidade de os mesmos conseguirem comunicar entre si (Gibaud, 2009). De forma a tornar isso possível, foram criados alguns *standards* para possibilitar a partilha de dados entre os diferentes sistemas, sendo os mesmos apresentados de seguida.

***Systematized Nomenclature of Medicine Clinical Terms* (SNOMED CT)**

SNOMED CT consiste num dicionário digital de termos médicos (conceitos, definições, relações, códigos, entre outros…) para utilização em contextos clínicos, especialmente em registos médicos digitais. A sua utilização visa a uniformização de termos nos diversos sistemas de informação hospitalares (Wang, Sable, & Spackman, 2002).

***Health Level 7* (HL7)**

HL7 é uma *framework* que é composta por um conjunto de *standards* associados, utilizados para permitir a interoperabilidade de registos médicos digitais, ou seja, para permitir a partilha de registos médicos entre diferentes sistemas de informação (HL7, 1987). Este *standard* especifica não só de que forma a informação deve ser armazenada, mas também como deve ser transmitida.

***Picture Archiving and Communication System* (PACS)**

O PACS trata-se de uma tecnologia que permite o armazenamento económico e transferência de imagens médicas. Todo o mecanismo associado a transferência das imagens é baseado na utilização de *queries* que permitem a realização de uma série de operações (como: aceder; mover; copiar; entre outras) sobre as imagens médicas (Choplin, Boehme, & Maynard, 1992). Por exemplo, no caso previamente apresentado do *workflow* de radiologia, se analisarmos a situação a um nível mais detalhado, após o equipamento efetuar a captura das imagens, as mesmas seriam enviadas para um servidor PACS, sendo posteriormente descarregadas pelo médico para a sua *workstation* de trabalho para serem analisadas.

***Digital Imaging and Communications in Medicine* (DICOM)**

DICOM é um *standard* para produzir, armazenar, apresentar, processar, partilhar e imprimir imagens médicas e outros documentos semelhantes (DICOM). É em tudo muito semelhante ao HL7, no entanto, DICOM é mais direcionado para a comunicação cliente-servidor, enquanto que o HL7 é mais utilizado na gestão de eventos.

1.4 Problema

Anteriormente foram apresentadas três formas amplamente utilizadas na elaboração de relatórios clínicos: *templates* estáticos; transcrição manual de uma gravação áudio para texto; e transcrição automática para texto. Qualquer um destes processos de elaboração de relatórios médicos podem implicar algumas limitações que são importantes de referir. A utilização de *templates* estáticos implicam um desperdício de tempo considerável por parte dos médicos na redação dos relatórios. A técnica baseada na transcrição manual do áudio gravado pelos médicos para documentos de texto, tarefa executada pelos técnicos de transcrição, traz custos extra consideráveis para o hospital na contratação desses recursos humanos. Quanto à transcrição automática, realizada por um sistema, implica também um aumento considerável de custos em *hardware/software*. Esta técnica também envolve um passo inicial de treino do sistema, que dura aproximadamente 30 minutos, onde cada médico deve ler uma série de frases pré-feitas que visam melhorar a eficácia do sistema no reconhecimento da fala. Qualquer uma destas técnicas também é propicia a erros ortográficos. Na técnica baseada em *templates* estáticos esses erros são originados pelo médico, enquanto que nas outras duas técnicas, baseadas na transcrição, os erros são provenientes de uma incorreta transcrição. Outro fator também limitativo, passa pela existência de médicos, em particular de faixas etárias mais velhas, que não se sentem confortáveis com novas tecnologias, ou mesmo com estes novos métodos de trabalho, preferindo muitas das vezes serem eles próprios a redigir os relatórios. Resumindo, identificam-se uma série de problemas que são comuns a muitos hospitais, sendo eles:

- O tempo gasto na redação e revisão dos relatórios (Rosenthal, et al., 1997);
- Os gastos associados aos recursos humanos e/ou *hardware/software* necessário;
- Erros ortográficos originados pelo médico ou por uma incorreta transcrição;
- Resistência na utilização das técnicas atuais por parte das faixas etárias mais velhas;

Face aos problemas expostos, identifica-se uma necessidade que passa por encontrar uma forma de agilizar o processo de elaboração de relatórios, que seja confortável, que reduza a quantidade de erros ortográficos e que, se possível, diminua os gastos associados a este processo. Com vista a resolver os problemas identificados, propõem-se o desenvolvimento de um sistema de previsão de texto que possa ser facilmente integrado num editor de texto. Este sistema será apresentado com maior detalhe no capítulo 4 e 5.

1.5 Objetivos

Como referido anteriormente, pretende-se desenvolver uma solução que permita aos médicos agilizar o processo de elaboração de relatórios clínicos, sugerindo as próximas palavras a serem utilizadas. Para tal, e como em qualquer projeto de engenharia, foram definidos uma série de objetivos a cumprir, sendo eles:

1. Aquisição e sintetização dos conhecimentos necessários;
2. Investigação do estado da arte de projetos similares;
3. Desenvolvimento de um modelo de previsão de palavras na elaboração de relatórios;
4. Desenvolvimento de um modelo de aprendizagem para incluir no modelo anterior;
5. Implementação dos modelos num protótipo;
6. Avaliação do protótipo.

Ao longo deste documento será apresentado ao leitor a forma como todos os objetivos foram concretizados, assim como os seus resultados.

1.6 Estrutura do relatório

No presente capítulo é feita uma breve apresentação desta dissertação, onde é explicado o seu âmbito, empresa onde foi realizada, todo o contexto envolvente, o problema identificado, objetivos e a estrutura deste documento. No capítulo 2 este trabalho é exposto numa ótica de negócio, abordando-se alguns conceitos relacionados com inovação, valor e modelo de negócios. Seguidamente, no capítulo 3 é apresentado um enquadramento teórico. Inicialmente são apresentadas as áreas que o trabalho desenvolvido abrange, depois são abordados alguns conceitos teóricos necessários para a compreensão deste documento e finalmente a investigação do estudo da arte efetuado, incluindo algumas aplicações relacionadas. No capítulo 4 são abordados tópicos relacionados com o *design* da solução a implementar, como as tecnologias utilizadas, uma análise das abordagens possíveis e a arquitetura do sistema. No capítulo 5 é exposto ao utilizador o processo de desenvolvimento da solução, incluindo os contratempos que surgiram. De seguida, no capítulo 6 é apresentado todo o processo de avaliação que vai ser utilizado para análise do desempenho do sistema desenvolvido, incluindo a definição das métricas a avaliar, a formulação das hipóteses a testar, resultados e respetiva interpretação. Finalmente, no último capítulo, 7, são apresentadas as conclusões da presente dissertação, onde se faz um resumo dos principais tópicos deste documento, abordam-se as limitações e trabalho futuro e faz-se uma apreciação final deste projeto.

2 Análise de valor

Neste capítulo será feita uma avaliação global do presente trabalho numa perspetiva de negócio. Numa primeira fase será apresentado um dos principais produtos onde o sistema a desenvolver poderá ser integrado, sendo de seguida apresentada a análise de valor da solução desenvolvida.

2.1 Produto alvo

Como já referido anteriormente, a FUJIFILM encontra-se a apostar no desenvolvimento de sistemas médicos, podendo o sistema desenvolvido, dependendo da sua avaliação, ser incorporado em outros produtos já comercializados. Um dos principais produtos desenvolvidos é o *Clinical Workflow Manager* (CWM), que é um grande sistema de gestão de imagens e informação médica. Este sistema consegue depois ser configurado para suportar diferentes sistemas de informação, como o de: radiologia (RIS); gastroenterologia (GIS); mamografia (BIS); entre outros (FUJIFILM). Um dos pontos mais fortes do CWM passa pela possibilidade de ser integrado com outros sistemas de informação dado ao facto de o mesmo suportar os vários protocolos anteriormente apresentados, uma mais valia para qualquer sistema de informação hospitalar. Relativamente à solução a desenvolver, a mesma poderá ser integrada, por exemplo, com o módulo de *reporting* do CWM, mais concretamente com a fase de elaboração de relatórios clínicos.

2.2 Análise de valor

Um dos principais conceitos inerentes a uma dissertação é o de inovação, sendo por isso importante compreender este termo. Inovação pode ser definido como uma mudança que trás valor, podendo este processo ser dividido em três partes: *Fuzzy Front End* (FFE); *New Product Development* (NPD); e comercialização (Koen, 2014); sendo neste momento interessante apresentar a primeira fase, FFE.

Fuzzy Front End é todo o processo desde o inicio de um projeto até ao inicio do seu desenvolvimento. De forma a que toda a gente consiga compreender este processo, foi criada uma linguagem comum que explica esta fase, sendo esse modelo conhecido por *New Concept Development model* (NCD). Uma esquematização deste conceito é apresentada na Figura 3.

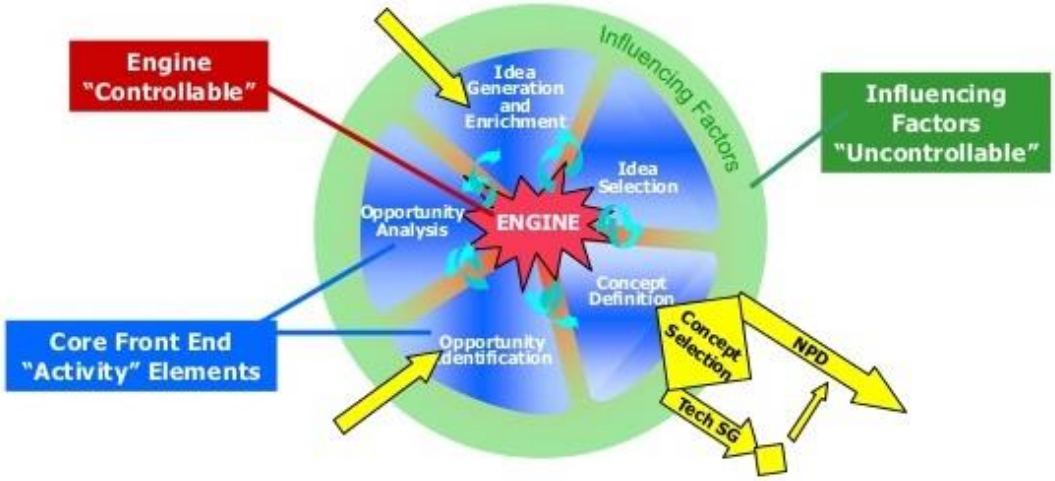

Figura 3 – *New Concept Development* (Koen, 2004).

Como se pode ver na imagem, o NCD pode ser dividido em três partes distintas, correspondentes a cada uma das três cores visíveis. A área vermelha é a responsável pela *front end*, a mesma corresponde ao grupo executivo que a controla. A área azul representa cinco atividades chave do processo de *front end.* E finalmente, a área verde corresponde aos fatores ambientais externos que podem influenciar as outras duas partes anteriormente apresentadas (Koen, 2014). Destas três secções do NCD, abordaremos as atividades chave com maior detalhe, sendo que as mesmas consistem nas seguintes cinco fases:

- Identificação de oportunidades;
- Análise de oportunidades;
- Geração e desenvolvimento de ideias;
- Seleção de ideias;
- E definição do conceito.

Alguns dos métodos e técnicas que podem ser aplicadas para análise destas cinco atividades chave durante a FFE são: Análise/*Benchmarking* das tecnologias existentes; planeamento de cenários; modelo das cinco forças de *Porter*; análise de mercados/concorrência; curvas S (Boretos, 2012); entre outros (Jauregi & Justel, 2007).

Relativamente ao caso especifico deste projeto, a identificação de oportunidades foi feita tendo em conta a evolução dos sistemas de informação hospitalares, onde se verificou que os mesmos tendem cada vez mais a basear-se na utilização de estruturas digitais. Um dos principais aspetos

positivos desta evolução, é a possibilidade da passagem imediata de informação entre os vários interessados, no entanto, o processo de elaboração de relatórios clínicos ainda se trata de uma fase demorada, tendo-se identificado uma necessidade que passa pela agilização desse processo. Após identificada a oportunidade, foi feita a sua análise, onde se verificou que uma ferramenta do género poderia ser uma mais valia se integrada numa série de produtos já comercializados para a área da saúde, tratando-se por isso de uma oportunidade a explorar. Foi então realizada uma sessão de *brainstorming* onde foram enumeradas várias ideias para por em prática. Depois, numa reunião com o gestor de projeto, as ideias sugeridas foram analisadas e desenvolvidas com mais profundidade, tendo-se no final escolhido uma ideia. Finalmente, definiu-se o conceito, tendo o mesmo sido apresentado numa feira da especialidade, onde recebeu *feedback* positivo. Com este sinal, foi iniciado o processo de planeamento e desenvolvimento.

2.2.1 Valor

O sistema a desenvolver durante a presente dissertação procura dar resposta a uma necessidade real e comum a muitos hospitais: o gasto excessivo de recursos (tempo e dinheiro) associado a elaboração de relatórios clínicos. Como tal, a solução a ser desenvolvida visa agilizar o processo de elaboração de relatórios médicos, sugerindo as próximas palavras a serem utilizadas pelo profissional de saúde. O sistema a desenvolver oferece uma série de vantagens tanto para o cliente (empresas prestadoras de cuidados de saúde), como para o utilizador final (médicos). Para as empresas permitirá diminuir os custos associados ao processo de elaboração de relatórios, isto através da eliminação da necessidade da contratação de transcritores, ou de *hardware*/*software* especifico para a transcrição. Relativamente ao utilizador final, o sistema proposto permitirá aos médicos poupar tempo na redação dos relatórios, podendo os mesmos focarem-se ainda mais na sua tarefa principal, a prestação de cuidados de saúde. Outra vantagem do sistema a desenvolver é que o mesmo é personalizável, adaptando-se automaticamente ao médico, o que implica que as previsões sejam mais exatas.

Como já anteriormente referido, a solução desenvolvida poderá ser integrada num produto já comercializado pela FUJIFILM, o que permitirá aumentar o valor das soluções. Em suma, o sistema desenvolvido permitirá que sejam os próprios médicos a redigirem os relatórios clínicos, de forma rápida e eficaz. Permitirá também que as empresas poupem dinheiro, dispensando a contratação de transcritores ou compra de *hardware/software* necessário.

2.2.2 Modelo Canvas

O modelo canvas trata-se de uma ferramenta que permite criar e apresentar de forma curta, sintética e visual um modelo de negócio, que consiste na forma como a empresa se estrutura para gerar valor (Kaminski & Enachev, 2014). De seguida são apresentadas cada uma das nove secções do modelo canvas, sendo o modelo em si apresentado na Figura 4.

Como já anteriormente referido, as atividades chave que a FUJIFILM desempenha em Portugal passam pelo desenvolvimento de soluções médicas (produtos e serviços) e respetivo suporte. Estas soluções são direcionadas para um segmento de mercado muito especifico, empresas prestadoras de cuidados de saúde, que podem ir desde hospitais centrais a pequenas clínicas especializadas (em radiologia, gastroenterologia, entre outros). Cada cliente tem um conjunto de necessidades muito especificas e muitas das vezes únicas, o que implica que as soluções desenvolvidas sejam muitas das vezes direcionadas unicamente para cada cliente. A execução destas atividades chave obriga à contratação de uma equipa qualificada para o desempenho das tarefas relacionadas com o desenvolvimento e suporte de *software*, mesmo assim, dado a enorme quantidade de projetos, foi necessário a criação de parcerias para ajudar a responder às necessidades dos vários clientes. Os parceiros chave com que a FUJIFILM conta são Seamlink e a Medsky, que ajudam com o desenvolvimento de *software* e *hardware*, respetivamente.

Relativamente ao valor das soluções oferecidas para os clientes, estas permitem em grande parte a conversão das infraestruturas baseadas em papel para uma infraestrutura digital, o que trás uma redução de custos, assim como dos riscos, associados aos registos em papel. As soluções desenvolvidas são também altamente customizáveis, podendo ser adaptadas aos vários serviços de um hospital.

Quanto às relações com clientes**,** estas podem ser consideradas relações de proximidade, com desenvolvimento de soluções vocacionadas para o cliente, onde é oferecido suporte 24/7 e atualizações. Existe ainda um portal onde os clientes podem consultar manuais, testar conhecimentos através de testes, contactar o suporte, e consultar novidades. A divulgação do produto é feita através do *site* oficial da empresa e feiras da especialidade.

Finalmente, os custos existentes estão associados ao arrendamento do espaço de trabalho e ainda ao pagamento dos salários dos funcionários (custos fixos). Como custos variáveis existe a compra esporádica de *hardware* e *software* necessário para a execução das atividades chave, e ainda a subcontratação de serviços, por exemplo, empresa de limpeza e *designers* gráficos. No lado oposto da balança, as principais fontes de receita advêm da venda de produtos e serviços, renovação de licenças, e ainda do suporte disponibilizado 24/7.

De seguida é apresentado o modelo canvas que resume a informação apresentada.

<table>
<tr>
<td rowspan="2">Parcerias Chave
 
Seamlink;
Medsky;</td>
<td>Atividades Chave
 
Desenvolvimento de soluções (produtos e serviços);
Suporte ao cliente;</td>
<td rowspan="2">Proposta de valor
 
Redução de custos;
Redução de risco;
Alta customização;</td>
<td>Relações com clientes
 
Relação de proximidade;
Suporte 24/7;</td>
<td rowspan="2">Segmentos de mercado
 
Empresas prestadoras de cuidados de saúde:
- Hospitais;
- Clínicas;</td>
</tr>
<tr>
<td>Recursos Chave
 
Mão de obra qualificada;
Equipamento médico proprietário;
Marca FUJIFILM;</td>
<td>Canais
 
Site oficial;
Feiras da especialidade;
Portal;</td>
</tr>
<tr>
<td colspan="3">Estrutura de custos
 
Salários e renda do espaço (custos fixos);
Outsourcing como limpeza, designers, etc… (custos variáveis);
Hardware/software para a realização das atividades chave (variáveis);</td>
<td colspan="2">Fontes de renda
 
Venda de soluções (produtos e serviços);
Renovação de licenças;
Suporte ao cliente;</td>
</tr>
</table>

Figura 4 – Modelo canvas.

3 Enquadramento teórico

Para uma melhor compreensão do trabalho desenvolvido, é necessário realizar todo um enquadramento teórico. Assim, o presente capítulo visa apresentar todo o conhecimento técnico necessário para a compreensão da solução implementada. Inicialmente serão indicadas as áreas abrangidas pelo nosso problema, sendo de seguida apresentados todos os conceitos teóricos abordados. No final, é apresentado o estado da arte e algumas soluções de previsão de texto existentes.

3.1 Áreas abrangidas

A solução proposta baseia-se no desenvolvimento de uma solução que agilize a elaboração de relatórios clínicos, através da previsão das próximas palavras a serem utilizadas. Este problema, o de previsão de texto, é conhecido na área de informática por *Language Modeling* (LM) e faz parte de um largo conjunto de problemas similares que pertencem ao ramo de Processamento de Linguagem Natural (PLN), responsável pela investigação e desenvolvimento de métodos/processos que permitam às máquinas compreender e manipular a linguagem natural das pessoas. Alguns dos problemas mais comuns que esta área visa resolver são: tradução de línguas; extração de informação em texto; reconhecimento de fala; identificação de erros gramaticais e de sintaxe; comunicação entre pessoas com capacidades reduzidas; entre outros (Nadkarni, Machado, & Chapman, 2011).

Como se não bastasse a complexidade do PLN em si, isto devido à grande diversidade de palavras, à ambiguidade, raridade, complexidade sintática, e diversos outros fatores (Manning & Schutze, 1999), o PLN encontra-se também em constante interação com outros ramos por si só já complexos: Inteligência Artificial (IA), que inclui *Machine Learning* (ML); Estatística e Probabilidade; Análise gramatical e lexical; *Data Mining*, mais concretamente, *Text Mining*; entre outros.

De seguida, são apresentados alguns conceitos teóricos destas áreas que são essenciais para a compreensão dos próximos capítulos.

3.2 Conceitos

A maioria dos sistemas que exibem comportamentos inteligentes baseiam-se na utilização de conhecimento para tomar decisões ponderadas. Esse conhecimento consiste na verdade num conjunto de dados processados até se tornarem em informações úteis, a que se dá o nome de base de conhecimento. A criação do conhecimento pode ser feita de forma manual, por um operador humano, ou então, de forma automática, por um sistema. Quando realizado de forma automática, o processo é conhecido por "treino" e obriga à utilização de dados já interpretados por uma pessoa. No contexto do presente trabalho, pretende-se desenvolver um sistema que exiba este comportamento, onde os dados interpretados para fornecer ao processo de treino consistirão em relatórios clínicos redigidos por médicos. Estes serão alvo de um processamento, devendo gerar no final a base de conhecimento que será utilizada para a previsão de texto. Todo este processo é resumido na figura 5.

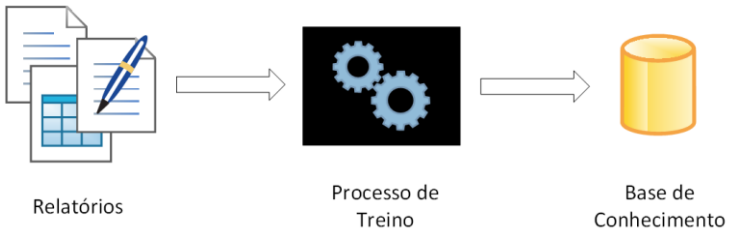

Figura 5 – Processo completo de treino (sem detalhes).

Alguns sistemas conseguem ainda aprender com situações que nunca ocorreram, preparando-se assim para situações futuras. Este processo de aprendizagem é geralmente conhecido na área de IA por *Machine Learning*. Novamente, é também pretendido que a solução desenvolvida exiba este comportamento, esquematizado na figura 6.

Figura 6 – Processo completo de aprendizagem (sem detalhes).

É importante deixar claro que os processos de treino e de aprendizagem são processos idênticos entre si (ou pelo menos quase idênticos). A diferença que se considera haver está unicamente relacionada com quando são executados. Consideramos que estamos perante um processo de treino quando o *input* é um conjunto de relatórios e não existe uma base de conhecimento, sendo a mesma criada do zero. Quanto ao processo de aprendizagem, consideramos o mesmo quando já existe uma base de conhecimento, e a mesma é alvo de uma atualização. No presente capitulo os detalhes do processo de treino e aprendizagem serão deixados de fora, sendo os mesmos apresentados com maior detalhe no capitulo 4.3.2.

Na área de *Language Modeling*, a um documento de texto, que no contexto do presente trabalho corresponde a um relatório, é dado o nome de *corpus*, ou *corpora* quando no plural (Copestake, 2004). Relativamente ao elemento atômico dum *corpus*, que pode ou não corresponder a uma palavra, é dado o nome de *token*. Por exemplo, a frase *"Olá, como estás?"* pode ser segmentada nos seguintes cinco *tokens*:

1) *Olá*
2) ,
3) c*omo*
4) *estás*
5) *?*

Dependendo do problema em questão, o conceito de *token* pode ser mais abrangente ou restrito. Num problema em que sinais de pontuação não sejam importantes, os mesmos podem ser retirados da definição de *token*, tornando assim o conceito mais restrito. Por outro lado, se numa outra situação forem importantes, estes podem ser incluídos na lista de *tokens* (Cambridge University, 2009).

Duas propriedades das palavras que são importantes conhecer, principalmente pelos problemas que colocam na área de PLN, são a variância e a ambiguidade. A variância ocorre quando existem várias palavras que podem ser usadas para fazer passar uma ideia, enquanto que a ambiguidade surge quando uma palavra pode ter mais que um sentido ou significado (Vallez & Jimenez, 2007). Estes dois problemas, de variância e ambiguidade, podem ser resolvidos tendo em conta o contexto de uma palavra, que corresponde ao conjunto de *tokens* que a antecedem.

Na previsão de texto deseja-se prever a próxima palavra tendo em conta o seu contexto. Tal pode ser calculado com recurso a *language models*, modelos que atribuem uma probabilidade a uma sequência de palavras. Por exemplo, utilizando o símbolo T para representar um *token,* poder-se-ia apresentar uma sequência de *N tokens* da seguinte maneira:

$$T_1, T_2, T_3, \dots, T_N \quad ou \quad T_1^N \tag{1}$$

Para indicar que T_N sucede $T_1, T_2, T_3, \dots, T_{N-1}$, é normal utilizar a seguinte nomenclatura:

$$T_N | T_1, T_2, T_3, \dots, T_{N-1} \quad ou \quad T_N | T_1^{N-1} \tag{2}$$

Considerando a equação 2, é depois possível modelar a probabilidade *P* do *token* T_N suceder $T_1, T_2, T_3, \dots, T_{N-1}$ da seguinte forma:

$$P(T_N \mid T_1, T_2, T_3, \dots, T_{N-1}) \quad ou \quad P(T_N \mid T_1^{N-1}) \tag{3}$$

As sequências analisadas podem depois ser decompostas em sequências mais pequenas, sendo a probabilidade final a mesma, esta propriedade matemática é conhecida como regra do produto. Aplicando este princípio à equação 3, temos:

$$P(T_N|T_1,T_2,T_3,\dots,T_{N-1}) = P(T_1)P(T_2|\,T_1)P(T_3|\,T_1,T_2)\dots P(T_N|T_1,T_2,T_3,\dots\,,T_{N-1})$$
$$= P(T_1^1)P(T_2|T_1^1)P(T_3|T_1^2)\dots\, P(T_N|T_1^{N-1}) \quad (4)$$
$$= \prod_{K=1}^{N} P(T_K|T_1^{K-1})$$

É importante realçar que, mesmo com a utilização de uma grande quantidade de documentos no processo de treino, é inexequível fornecer ao sistema todas as sequências de *tokens* possíveis, tornando-se assim natural que algumas das probabilidades decompostas apresentadas na equação 4, sejam impossíveis de ser calculadas. De forma a dar a volta a este problema, pode-se tentar aumentar a probabilidade de encontrar uma sequência de *tokens* utilizando um contexto mais pequeno. Isto implica assumir que o contexto mais distante de uma palavra não influência a previsão dos próximos *tokens*, assunção essa a que é dada o nome de *Markov assumption*. Aplicando este princípio sobre a equação 4 e sendo *C* o tamanho do contexto, tem-se:

$$P(T_N|T_1,T_2,T_3,\dots,T_{N-1}) \approx \prod_{K=N-C}^{N} P(T_K|T_1^{K-1}) \quad (5)$$

Finalmente, o último passo necessário compreender é o calculo de cada probabilidade decomposta, que é determinada tendo em conta a frequência *F* das respetivas sequências de *tokens*. Por exemplo, para a equação 5, o valor da última probabilidade decomposta, quando *K = N*, poderia ser calculada da seguinte forma:

$$P(T_N|T_1^{N-1}) = \frac{F(T_N|T_1^{N-1})}{F(T_1^{N-1})} \quad (6)$$

Concluindo, para prever a próxima palavra tendo em conta uma frase redigida por um utilizador, é necessário em primeiro lugar identificar todos os *tokens* que sucedem essa sequência. De seguida, considerando as equações 5 e 6, é possível calcular as probabilidades de cada *token* possível, e assim, apresentar ao utilizador os mais prováveis.

No capítulo 3.3, estado da arte, serão abordados os métodos mais comuns utilizados na previsão de texto. Desses, destacam-se três, que serão apresentados de seguida com maior detalhe, sendo eles: *NGrams*; Redes Neuronais (RN); e *Suffix Trees*.

Os *NGrams* são um modelo probabilístico que agrupa um documento em sequências de *N tokens*. Tendo em conta os conceitos até agora apresentados (Jurafsky & Martin, 1999), é possível utilizar estes modelos para a previsão de texto. Por exemplo, aplicando um *NGram* de ordem 2 na frase "*Olá, como estás?*" teríamos os seguintes *NGrams*:

- Olá ,
- , como
- como estás
- estás ?

Seria depois possível calcular P(,|Olá), P(como|,), … e assim, estimar o próximo *token* mais provável.

Outra técnica também passível de ser utilizada são as redes neuronais que, procuram imitar o funcionamento dos neurónios dos seres vivos, principalmente no aspeto relacionado com a capacidade de processamento paralelo de informação. As RN também necessitam de um processo de treino para gerar a sua rede de neurónios e têm como base alguns dos conceitos anteriormente apresentados. Uma das maiores vantagens das RN é o facto de não obrigarem à utilização da *Markov assumption*, ou seja, possibilitam a utilização de contextos muitos longos para a previsão de *tokens* (Kriesel, 2005). Na Figura 7 é apresentada uma esquematização básica de uma rede neuronal para a previsão de texto, onde $T_1, T_2, T_3, \dots, T_N$ são os *tokens* de contexto, e T_{N+1} o *token* previsto.

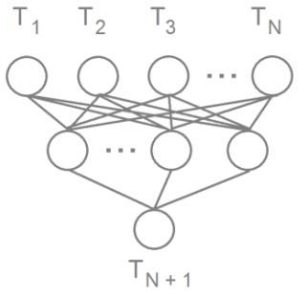

Figura 7 – RN para previsão de texto.

A camada superior, conhecida por *input layer*, deverá receber os *tokens* que pertencem ao contexto da palavra que se deseja prever. A camada central, *hidden layer*, que na maioria das situações é constituída por mais que uma camada, procura transformar o *input* em *output* através da aplicação de uma série de funções. Por último, a última camada, *output layer*, corresponde ao *token* previsto.

Finalmente, a última técnica que será posteriormente abordada são as *Suffix Trees*, que são estruturas de dados em forma de árvore amplamente utilizadas para operações com *strings*. A vantagem destas estruturas é a sua velocidade de pesquisa, no entanto, para poder efetuar essas pesquisas rapidamente, é necessário um pré processamento muito demorado, sacrificando assim a velocidade durante o processo de treino (Kay). Geralmente, nesta técnica cada nó representa um caractere, no entanto, a mesma pode ser adaptada para que cada nó contenha um *token*.

3.3 Estado da arte

Será agora apresentado ao leitor a investigação do estado da arte realizada no âmbito da presente dissertação. Inicialmente serão apresentados os primeiros trabalhos publicados nas áreas de IA e PLN, partindo depois para o ramo mais concreto de LM.

3.3.1 Origem

Como já referido anteriormente, a solução proposta é baseada na previsão de texto, previsão essa que pode ser realizada com recurso a métodos de Processamento de Linguagem Natural, uma área intimamente ligada à Inteligência Artificial. O conceito de IA surgiu em 1950, quando Alan Turing

publicou um artigo com o título "*Computer Machinery and Inteligence*". Ao longo do artigo, Turing aborda a questão "*Can machines think?*", introduzindo assim o conceito de Inteligência Artificial (Turing, 1950) e iniciando uma nova era no mundo dos computadores. Com esta primeira pegada de Turing, muitos outros autores começaram a desenvolver e publicar trabalhos sobre a área. Uns anos mais tarde, em 1954, a *International Business Machines* (IBM) apresenta o primeiro sistema de PLN. Este sistema consistia num computador que traduzia textos redigidos em russo para inglês, utilizando um dicionário de termos traduzidos composto por 250 palavras e na codificação de seis regras gramaticais (Hutchins, 2005). Um dos maiores problemas deste sistema eram as palavras ambíguas que, sem o seu contexto, poderiam ser mal traduzidas (IBM, 1954).

3.3.2 Atualidade

Atualmente, problemas de PLN podem ser resolvidos com recurso a três tipos de abordagens, sendo elas: abordagens baseadas em regras; abordagens baseadas em estatística e probabilidades; e abordagens híbridas, que junta aspetos de ambas.

3.3.2.1 Abordagens baseadas em regras

As abordagens baseadas em regras assentam, como o próprio nome indica, num conjunto de regras que são utilizadas pelo sistema para tomar uma decisão relativamente a um *input*. Este conjunto de regras representam a base de conhecimento do domínio e são codificadas durante a fase de desenvolvimento do projeto. Vitoria e Abascal desenvolveram um sistema de previsão de palavras baseada nesta abordagem (Vitoria & Abascal, 1997). Este consistia na análise sintática de texto, onde inicialmente era realizado um *parse* sintático da frase e, de seguida, a partir de um conjunto de regras sintáticas previamente codificadas, era feita a previsão da próxima palavra dessa frase. O sistema da IBM (IBM, 1954), anteriormente referido, também segue esta abordagem baseada em regras estáticas.

Uma das maiores vantagens destas técnicas é a sua simplicidade de implementação, principalmente pelo facto de se tratarem de um método primariamente declarativo, tornando-se assim de fácil compreensão e manutenção. A incorporação de novo conhecimento de domínio trata-se de uma tarefa simples, sendo apenas necessário adicionar a regra criada à base de conhecimento. No entanto, esta tarefa apenas pode ser realizada manualmente e durante a fase de desenvolvimento. Quanto à criação das regras, tal processo exige um enorme esforço manual e o aumento da quantidade das mesmas implica também o aumento da complexidade do sistema. De seguida, é apresentada uma tabela com o resumo dos principais aspetos destas técnicas.

Tabela 1 – Resumo das abordagens baseadas em regras.

Aspetos chave
Implementação simples
Fácil compreensão e manutenção
Fácil resolução de erros
Processo de incorporação de novo conhecimento simples, mas manual
Criação de regras exige um enorme esforço manual
Complexidade aumenta com o número de regras

De todos os aspetos apresentados, destaca-se a incapacidade destes sistemas aprenderem sozinhos, sendo necessário um operador humano para a incorporação de novo conhecimento. Devido a este fator, esta abordagem é considerada obsoleta no meio académico, sendo deixada de fora ou, pelo menos, ofuscada em muitos trabalhos de investigação. A nível comercial ocorre o oposto, existindo maior ênfase na utilização de regras. Tal deve-se ao facto de as pessoas desconfiarem da capacidade da IA, sendo mais seguro a nível comercial a venda e divulgação de produtos baseados em regras (Chiticariu, Li, & Reiss, 2013).

3.3.2.2 Abordagens baseadas em estatística e probabilidades

Aproximadamente ao mesmo tempo, começaram a surgir sistemas de PLN baseados em estatística e probabilidade. Estes sistemas procuram colmatar alguns dos maiores problemas da técnica anterior, como a incapacidade de aprendizagem de forma automática. Abascal e Vitoria, já anteriormente referenciados, exploraram esta técnica tendo desenvolvido um sistema de previsão baseado na frequência das palavras a prever (Abascal & Vitoria, 1994). Outro trabalho similar, mas tendo em conta a data da última utilização da palavra anterior, foi publicado por outro grupo de investigadores em 1987 (Swiffin, Arnott, Pickering, & Newell, 1987). Deve-se realçar que ambos os trabalhos tiveram resultados negativos, principalmente por não terem em conta o contexto da palavra, um fator importante na linguagem natural.

A principal vantagem desta técnica em relação à abordagem baseada em regras, é a facilidade da adaptação ao meio ambiente, tratando-se de um método passível de ser treinado e adaptado automaticamente. Relativamente aos aspetos negativos, destacam-se dois: a complexidade destas soluções, muito superior à dos sistemas baseados somente em regras; e também a obrigatoriedade da realização de um treino prévio, que pode ser um passo demorado e que geralmente implica um consumo considerável de recursos computacionais. De seguida é apresentada uma tabela com o resumo dos principais aspetos desta abordagem.

Tabela 2 – Resumo das abordagens baseadas em estatística e probabilidades.

Aspetos chave
Implementação complexa
Adapta-se ao meio
Processo de incorporação de novo conhecimento complexa, mas automática
Reduz o esforço manual
Obriga a uma fase de treino que pode ser demorada
Requer dados interpretados para o processo de treino

Relativamente à área de *Language Modeling*, ramo responsável pelos problemas relacionados com a previsão de texto, esta pode utilizar ambas as abordagens apresentadas para a resolução de problemas associados, mas geralmente apresenta melhores resultados com as técnicas baseadas em estatística. Quando se fala sobre previsão de texto, pode-se ter em conta dois tipos de previsão, sendo eles, a da palavra e a da frase. Dos dois, o tipo de previsão que é interessante para este trabalho é o da frase.

Para além das soluções de previsão de texto já apresentadas, foram identificados outros métodos para resolver este problema, sendo um desses métodos baseado nos *NGrams*. *Trinh, Waller et all*

desenvolveram um sistema para comunicação de pessoas com deficiências físicas e de fala baseado na utilização de *NGrams* de ordem 6. Para medir os resultados do sistema implementado, foram utilizados os conceitos de *Keystroke Savings* (KS) e *Keystroke per Character* (KSPC), conceitos apresentados com detalhe no capítulo 6, tendo obtido valores de 56.3% e 19.1% respetivamente (Trinh, Waller, Vertanen, Kristensson, & Hanson, 2012). *Bickel, Haider e Scheffer* criaram também eles um sistema baseado em *NGrams* de ordem 2 e 3. Os resultados obtidos apresentaram um valor de KS entre os 2% e os 60%. Os autores chegaram a conclusão que a quantidade de dados no treino, assim como o tamanho do contexto e ainda a ordem dos *NGrams* utilizados são fatores importantes a ter em conta na previsão de texto, influenciando muito os resultados obtidos (Bickel, Haider, & Scheffer, 2005). Outras soluções encontradas utilizavam Redes Neuronais para a previsão de texto, (Bengio, Ducharme, Vincent, & Jauvin, 2003) e (Kannan, et al., 2016). Os dois trabalhos apresentaram bons resultados, mas a maior conclusão a que ambos chegaram foi que as RN permitiam que a previsão de texto tivesse em conta um maior número de palavras no contexto. Finalmente, outra abordagem encontrada baseava-se na utilização de *Suffix Trees*, este trabalho apresentou um valor de KS próximo dos 20% (Nandi & Jagadish, 2007).

No presente subcapítulo apenas se enumeraram as várias abordagens possíveis para a previsão de texto. A solução desenvolvida basear-se-á num destes métodos, no entanto, a escolha da técnica a utilizar apenas é apresentada no capítulo 4.3, onde é realizada uma análise entre as várias soluções identificadas.

3.4 Aplicações de previsão de texto

Ferramentas de previsão de texto podem ser uma mais valia na área das comunicações. Nos dias de hoje, estas ferramentas são amplamente utilizadas em dois cenários distintos e recorrentes. O primeiro desses cenários está relacionado com área de comunicações móveis. Com o crescimento exponencial da utilização do *Short Message Service* (SMS), as pessoas começaram a sentir a necessidade de escreverem cada vez mais depressa, sendo este um bom ponto de entrada para a previsão de texto. De forma a aumentar a velocidade de escrita das mensagens, sistemas de entrada de texto que facilitem esse processo por parte do utilizador têm vindo a ser desenvolvidos (Isokoski, 2004). Uma dessas formas passa pela previsão de texto que, na maioria dos sistemas que a implementa, consiste apenas na previsão da palavra a ser redigida. Hoje em dia, a esmagadora maioria dos dispositivos móveis, senão todos, são vendidos com um dicionário incorporado que permite essa previsão. Alguns desses sistemas são o *QuickType* (Filipowicz, 2016) implementado pela Apple e o *SwiftKey* (Henry, 2014) da Android. No entanto, quando se fala na previsão da frase, estes sistemas são muito limitativos, não se encontrando resultados publicados relativos a este tema. Tal pode ser explicado pelas limitações computacionais dos telemóveis: memória, capacidade de processamento e limitado consumo de energia, assim como ao enorme contexto temático das conversas por SMS, que dificulta a previsão (Durão, Rocha, & Carvalho, 2003). O reduzido número de teclas no teclado implica também que os utilizadores tenham de carregar em mais teclas para digitar um caractere, o que implica um elevado valor de KSPC.

O outro cenário onde a previsão de texto pode ser de grande utilidade é na comunicação de pessoas com problemas de fala. Sistemas desenvolvidos para este fim têm geralmente em conta dois fatores importantes, o algoritmo de previsão de texto e a ergonomia do teclado. Estes dois aspetos, quando

bem implementados, conseguem reduzir o esforço do utilizador, o que é ideal nestes sistemas. Na Figura 8 é apresentado o *layout* de um teclado preditivo desenvolvido para pessoas com deficiências que afetem a sua capacidade de comunicação. Como é possível ver na figura, estes teclados apresentam ao utilizador um conjunto de palavras que são utilizadas com uma elevada frequência no quotidiano, mas também possibilitam a escrita livre. Relativamente à previsão, estes sistemas possuem uma grande vantagem em relação a um teclado normal de telemóvel, que passa pela possibilidade de definir alguns contextos da conversa, o que pode melhorar os resultados da previsão (TouchChat). Apesar disso, o número de contextos oferecidos são limitados, não conseguindo cobrir todas as necessidades de uma conversa casual.

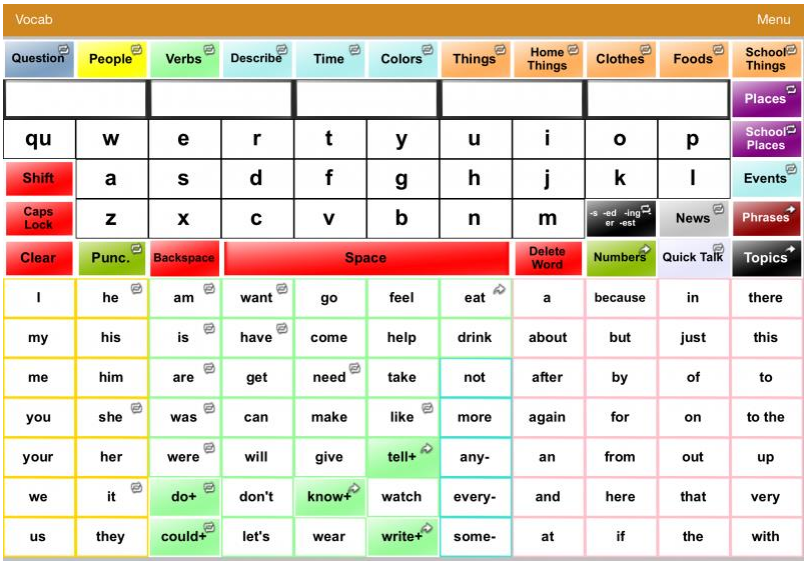

Figura 8 – Teclado preditivo da VocabPC (TouchChat).

No geral, verifica-se a existência de uma grande oferta de sistemas que facilitam a comunicação através da previsão de texto. No entanto, essa previsão é muito limitada, sendo apenas prevista a palavra corrente, principalmente devido a dificuldade de definir um contexto para a conversa. Um sistema que ultrapasse esta limitação, poderá efetivamente obter bons resultados, e será certamente uma mais valia para as pessoas.

4 *Design*

Neste capítulo é apresentado o planeamento e decisões sobre a arquitetura da solução a desenvolver. Inicialmente serão apresentadas as tecnologias utilizadas. De seguida, é utilizada uma ferramenta de tomada de decisão para a escolha da abordagem a seguir. No final, é apresentado o *design* e a arquitetura do sistema.

4.1 Tecnologias

A nível de tecnologias existiu uma grande influência do ambiente empresarial, tendo-se adotado inúmeras práticas da FUJIFILM. O ambiente de desenvolvimento utilizado foi o *Visual Studio 2015* (VS15) e a linguagem de programação o C#. Utilizou-se também um sistema para gestão do código fonte, *Team Foundation Server* (TFS) que, juntamente com o *Team Foundation Version Control* permitiu o controlo de versões.

4.2 Análise de possíveis soluções

Face o estudo do estado da arte efetuado no capítulo 3.3, foram identificados quatro métodos distintos para realizar a previsão de texto, sendo eles através da utilização de: regras; *NGrams*; redes neuronais; ou *Suffix Trees*.

Uma abordagem baseada em regras é bastante simples de implementar, no entanto, sofre por não suportar aprendizagem automática (AA), exigindo também um grande esfoço manual na elaboração das regras. Quanto aos *NGrams*, estes são uma abordagem amplamente utilizada e que oferece bons resultados. A sua implementação é um pouco mais complexa que a das regras, no entanto, continua a ser simples. Os *NGrams* também suportam aprendizagem automática e obrigam à utilização de dados previamente interpretados. Relativamente às RN, a sua implementação é a mais complexa das quatro abordagens, quanto aos outros parâmetros, aprendizagem automática e necessidade de dados já interpretados, as RN são idênticas aos *NGrams*. Uma das maiores vantagens das RN relativamente aos *NGrams* é que estas permitem ter em conta um contexto maior da frase para a previsão. No entanto, a assunção de *Markov* diz que contextos muito grandes não são importantes para uma

previsão, não sendo por isso este fator muito importante. Finalmente, as *Suffix Trees* apresentam uma complexidade de implementação média e ainda suportam aprendizagem automática. No entanto, tendo em conta a investigação do estado da arte efetuado, os resultados com esta abordagem não foram muito animadores. Na Tabela 3 é apresentado um quadro resumo com os principais aspetos de cada uma das abordagens possíveis.

Tabela 3 – Comparação de diversos fatores entre as várias abordagens possíveis.

	Regras	*NGrams*	RN	*Suffix Trees*
Complexidade de implementação	Simples	Média	Alta	Média
Resultados	Médios	Bons	Bons	Médios
Suporta aprendizagem automática	Não	Sim	Sim	Sim
Requer dados interpretados	Não	Sim	Sim	Sim

Para além dos aspetos apresentados, existem outros fatores que devem ser tidos em consideração ao escolher a abordagem a seguir.

- Grande quantidade de documentos disponíveis (120 000 relatórios clínicos);
- Contexto muito específico (radiologia, gastroenterologia, entre outros);
- Possibilidade de dividir os relatórios por autor.

De forma a auxiliar uma tomada de decisão correta relativamente a abordagem a seguir, foi utilizada uma ferramenta de tomada de decisão conhecida por *Analytic Hierarchy Process* (AHP). Esta ferramenta permite tomar uma decisão tendo em conta uma série de critérios qualitativos e quantitativos e pode ser dividida nas seguintes seis fases (Silva & Belderrain, 2005):

1 - Construção da árvore hierárquica de decisão

Esta fase consiste na modelação e estruturação do problema em forma de árvore de decisão, onde se realça: o objetivo final; os critérios associados; e as alternativas adequadas. No contexto da presente dissertação e tendo em conta o problema identificado, a árvore criada será a seguinte:

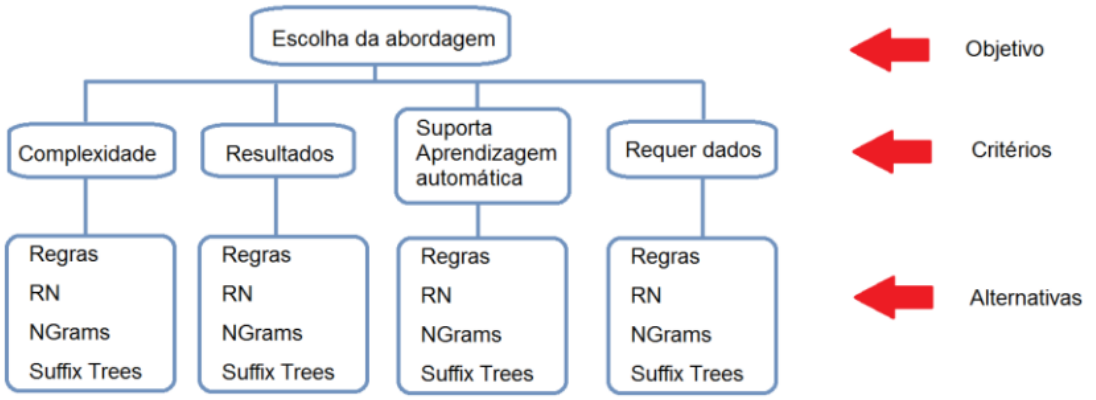

Figura 9 – Árvore de decisão.

2 - Comparação das alternativas e critérios

Nesta etapa são estabelecidos os níveis de comparação entre os diferentes critérios definidos no passo anterior. Para tal, foi utilizada a escala fundamental de Saaty (Saaty, 2008), que define os seguintes valores:

Tabela 4 – Escala fundamental de Saaty (Saaty, 2008).

Valor	Definição	Descrição
1	Igual importância	Os dois critérios contribuem de forma idêntica para os objetivos
3	Pouco mais importante	A análise e experiência mostram que um critério é um pouco mais importante que o outro
5	Muito mais importante	A análise e experiência mostram que um critério é claramente mais importante que o outro
7	Bastante mais importante	A análise e experiência mostram que um critério é predominante para o outro
9	Extremamente mais importante	Sem qualquer dúvida um dos critérios é absolutamente predominante para o objetivo
2, 4, 6, 8	Valores intermédios	Valores intermédios entre os níveis previamente apresentados

Considerando uma tabela 4x4 com os critérios definidos, é possível atribuir um valor a cada relação entre critérios, obtendo-se a seguinte tabela:

Tabela 5 – Comparação de critérios.

	Complexidade	Resultados	Suporta AA	Requer dados
Complexidade	1	1/5	1/5	1/3
Resultados	5	1	1/3	3
Suporta AA	5	3	1	3
Requer dados	3	1/3	1/3	1

3 - Calculo da prioridade relativa de cada critério

De seguida, tendo em conta a tabela 5, foram normalizados todos os fatores apresentados de forma a utilizar a mesma unidade para todos eles. Para a normalização, cada elemento da tabela deve ser dividido pelo total da respetiva coluna. Depois, calculando a média de cada linha é possível obter a prioridade relativa de cada critério, ou seja, para a tabela anterior temos agora:

Tabela 6 – Comparação entre critérios normalizada e prioridade relativa.

	Complexidade	Resultados	Suporta AA	Requer dados	Prioridade Relativa
Complexidade	0.07	0.04	0.11	0.05	0.07
Resultados	0.36	0.22	0.18	0.41	0.29
Suporta AA	0.36	0.66	0.54	0.41	0.49
Requer dados	0.21	0.07	0.18	0.14	0.15

Obtendo-se o vetor de prioridade *{0.07 0.29 0.49 0.15}*.

4 - Avaliação da consistência das prioridades relativas

Nesta fase é calculada a Razão de Consistência (RC) que, nos permite saber se os julgamentos feitos foram consistentes em relação a grandes amostras de juízos completamente aleatórios. Para tal, numa primeira fase deve ser calculado o valor próprio (Υ_{max}) que, considerando a equação 7, onde α é o vetor de prioridade relativa calculado na etapa 3 e *M* é a matriz 4x4 com os fatores de comparação da etapa 2, obtemos o Υ_{max} igual a *4.25*.

$$\alpha\Upsilon_{max} = \alpha M \tag{7}$$

Depois, é possível calcular o Índice de Consistência (IC) com a equação 8, onde *n* corresponde ao número de critérios, obtendo-se o IC de *0.08*.

$$IC = \frac{\Upsilon_{max} - n}{n - 1} \tag{8}$$

Finalmente, é possível calcular o RC considerando a equação 9, onde o IC foi calculado anteriormente tendo-se obtido o valor de *0.08*, e o índice aleatório (IR), é um valor tabelado que para matrizes 4x4 corresponde a *0.9*. Com isto, obtêm-se assim uma razão de consistência de *0.09*.

$$RC = \frac{IC}{IR} \tag{9}$$

Tendo em conta que o RC obtido é menor que *0.1*, podemos concluir que os valores das prioridades relativas estão consistentes, sendo assim possível prosseguir com o processo de decisão.

5 - Construção da matriz de comparação paritária para cada critério

Nesta etapa são repetidos todos os procedimentos para a construção da matriz de comparação e calculo da prioridade relativa para cada critério em cada uma das alternativas enumeradas (cálculos apresentados no Anexo A – Cálculos intermédios AHP). No final, para cada alternativa obtemos os seguintes vetores de prioridade:

	Regras	NGrams	RN	Suffix Trees
Complexidade	(0.52	0.20	0.08	0.20)
Resultados	(0	0.33	0.33	0.33)
Suporta AA	(0	0.33	0.33	0.33)
Requer dados	(0.10	0.28	0.47	0.16)

O que permite a construção da seguinte matriz:

$$M = \begin{pmatrix} 0.52 & 0 & 0 & 0.10 \\ 0.20 & 0.33 & 0.33 & 0.28 \\ 0.08 & 0.33 & 0.33 & 0.47 \\ 0.20 & 0.33 & 0.33 & 0.16 \end{pmatrix}$$

6 - Calculo da prioridade composta para as alternativas

Finalmente, é possível calcular as prioridades compostas para cada alternativa, multiplicando a matriz anterior pelo vetor de prioridades obtido no passo 3. O resultado final é o vetor de prioridade *{ 0.05 0.31 0.33 0.30 }* para as abordagens baseadas em regras, RN, *NGrams* e *Suffix Trees*, respetivamente.

Face todos os fatores apresentados na tabela 3, assim como os aspetos enumerados, e ainda suportando a decisão com a análise AHP realizada, conclui-se que as melhores abordagens para resolver o problema proposto baseiam-se na utilização de *NGrams* ou RN, isto porque apresentam as maiores prioridades no vetor final calculado. Ambas as técnicas suportam aprendizagem automática e apresentam bons resultados. Relativamente ao facto de necessitarem de dados já interpretados, tal não é um problema devido à disponibilidade de acesso a base de dados com milhares de relatórios clínicos. No geral tratam-se de soluções com pontos fortes e fracos muito similares, sendo a única diferença clara a complexidade de implementação. Devido a este aspeto, a abordagem a adotar será baseada nos *NGrams*.

4.3 *Design*

A fase de desenho do sistema pode ser dividida em duas fases. A primeira consiste na definição dos requisitos funcionais (RF) e não funcionais (RNF) do sistema. A segunda passa pelo desenho em si da solução, que inclui a arquitetura e o todos os *workflows* necessários. De seguida apresentamos ambas as fases.

4.3.1 Requisitos

Face aos problemas apresentados no capítulo 1.4, como: o tempo gasto na redação e revisão de relatórios clínicos (nos *templates* estáticos); gastos associados à contratação de transcritores para a transcrição manual; gastos associados à compra de *hardware*/*software* necessário para a transcrição automática; erros ortográficos (em qualquer abordagem possível); e desconforto na utilização destas técnicas por parte das faixas etárias mais velhas (tanto na transcrição manual, como na automática); pretende-se desenvolver um sistema que permita agilizar o processo de elaboração de relatórios. Para tal, sugere-se o desenvolvimento de uma ferramenta que, possa ser integrada num editor de texto e apresente ao utilizador sugestões de frases tendo em conta o que o mesmo redigiu. Um sistema deste género permitiria ao utilizador a elaboração de um relatório completo com pouco esforço, bastando-lhe escrever as primeiras palavras do relatório e ir selecionando as sugestões apresentadas pelo sistema. No caso de nenhuma das sugestões apresentadas ser a desejada pelo utilizador, bastar-lhe-ia escrever mais uma palavra para novas sugestões serem apresentadas. Um sistema deste género permitiria ao utilizador: poupar tempo na redação dos relatórios; eliminar a necessidade de contratação de pessoas para a transcrição manual, ou a necessidade da compra de *hardware*/*software* para a transcrição automática; reduzir os erros ortográficos (partindo do principio que a base de

conhecimento não apresenta esses erros); e permitir a redução do esforço do utilizador no processo de elaboração de relatórios clínicos, mas deixando todo o controlo deste processo no lado do mesmo.

Numa primeira fase deste projeto será desenvolvido um protótipo em forma de editor de texto que será posteriormente avaliado, este processo de avaliação será apresentado com detalhe no capítulo 6. A primeira tarefa a executar foi o levantamento de requisitos funcionais do projeto, levantamento esse que é apresentado de seguida na forma de um diagrama de *use cases*.

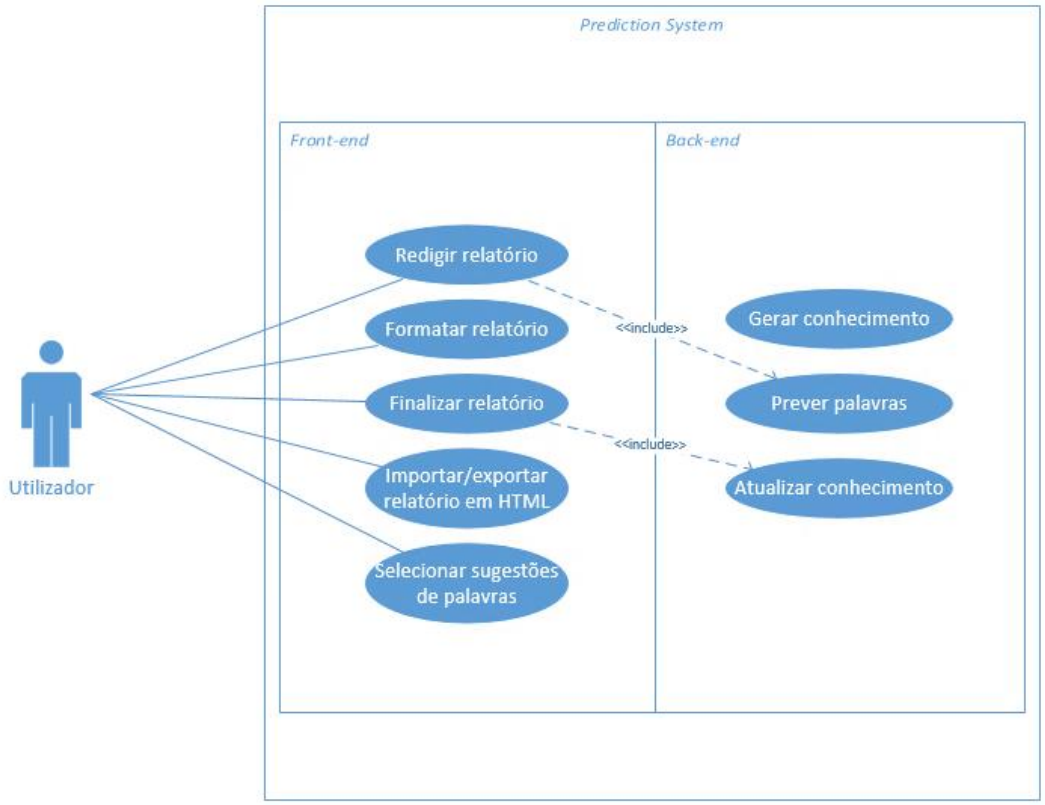

Figura 10 – Diagrama de *Use Cases*.

Com a definição dos RF foi possível dividir o sistema a desenvolver em dois módulos: *front-end* e *back-end*. O módulo de *front-end* será a *interface* que ligará o sistema em si ao utilizador. Como já referido, o mesmo terá a forma de um editor de texto básico que deverá permitir ao utilizador:

- A redação de relatórios clínicos;
- Aplicação de algumas formatações básicas nos relatórios;
- Finalização de relatórios;
- Importação e exportação de relatórios em HTML;
- Permitir que o utilizador selecione sugestões de palavras feitas pelo sistema.

Relativamente ao módulo de *back-end*, o mesmo consistirá no *core* do sistema e deverá:

- Gerar conhecimento a partir de uma base de dados com relatórios médicos;
- Prever as próximas palavras que o utilizador pode escolher;
- Atualizar a base de conhecimento quando um relatório é finalizado.

De seguida, foram definidos os requisitos não funcionais do projeto, onde foi dado enfâse à usabilidade e ainda a algumas questões legais do sistema. Em suma, o sistema deverá:

- Oferecer uma *interface* gráfica intuitiva, ergonómica e responsiva; (Usabilidade)
- Ter um tempo útil de resposta com sugestões inferior a 0.5 segundos; (Usabilidade)
- Filtrar todas as informações pessoais dos utentes ou médicos, como nomes e IDs, das sugestões oferecidas. (Legal)

Com os requisitos definidos, foi possível e mais fácil iniciar o planeamento e design do sistema a desenvolver, processo esse apresentado de seguida.

4.3.2 Arquitetura

Pretende-se ver desenvolvido um sistema que permita a previsão de texto mediante o *input* do utilizador. Existem alguns fatores já anteriormente abordados que devem ser relembrados, principalmente dada a sua importância e possível influência no desempenho do sistema, sendo eles:

- Ambiente altamente contextualizável;
- Possível divisão de relatórios por autor;
- Quantidade de documentos disponíveis.

Uma das maiores vantagens da solução a desenvolver, é que a mesma ocorre num ambiente altamente contextualizado, ou seja, o tema dos relatórios redigidos é sempre relacionado com medicina. É ainda possível ir a um nível de detalhe maior, sendo possível detalhar o tema ao nível de uma especialidade médica, como por exemplo, radiologia, gastroenterologia, mamografia, entre outros. Além do tema dos relatórios, é ainda possível realizar uma filtragem dos mesmos por autor. Apesar dos termos e conceitos utilizados pelos médicos serem os mesmos, muitas das vezes eles podem descrever a mesma ideia de forma diferente. As principais diferenças que se verificam estão relacionadas com a utilização dos verbos, quer seja pela utilização de sinónimos, como pela utilização de tempos diferentes, por exemplo, as frases “Verificam-se evidências de…”, “Verifica-se evidências de…” e “Notam-se evidências de…”, são exemplos de três frases diferentes que querem fazer passar a mesma ideia. Devido a esta variância, a possibilidade de filtrar os relatórios por autor é uma mais valia, visto que por norma um médico tende sempre a utilizar o mesmo estilo na redação dos relatórios. Outro fator também vantajoso para a solução, é a quantidade de dados disponíveis para treino. Qualquer hospital que deseje implementar este sistema, terá certamente milhares de relatórios guardados dos seus profissionais de saúde, sendo por isso relativamente fácil a criação duma base de conhecimento especifica para cada utilizador.

Quanto ao sistema em si, como já anteriormente referido, o mesmo pode ser dividido em dois módulos, o *back-end* e o *front-end*. O módulo principal do sistema será o *back-end* que, terá como principais tarefas a realização do treino, aprendizagem e previsão de texto. Se após a avaliação os resultados obtidos forem positivos, o sistema desenvolvido poderá ser utilizado em qualquer produto onde a previsão de texto possa ser uma mais valia, pretendendo-se assim o desenvolvimento de uma solução modular que, possa ser facilmente integrada noutras soluções. Para tal, decidiu-se que a mesma seria desenvolvida como uma *Dynamic Link Library* (DLL), que consiste numa coleção de código e dados que pode ser importada para qualquer projeto que necessite de tais funcionalidades. Por exemplo, no caso do CWM, apresentado no capítulo 2.1, o módulo responsável pela elaboração dos relatórios é o CWM *Reporting*, que inclui um editor de texto próprio. No caso da integração com o CWM, bastaria alterar o módulo de *reporting* para este incluir o sistema de previsão de texto. O módulo de *reporting* seria depois responsável por fornecer ao sistema o *input* do utilizador e apresentar as sugestões das previsões realizadas. Na figura 11 é apresentada uma esquematização deste processo.

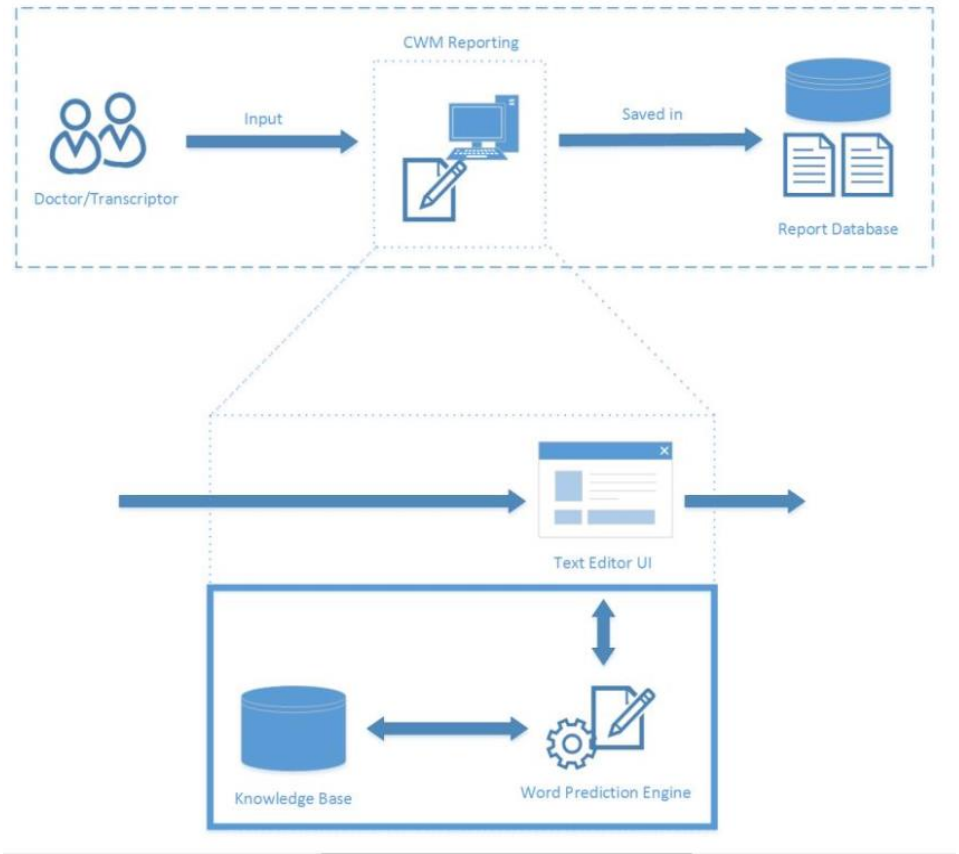

Figura 11 – *Flow* do CWM *Reporting* com o sistema de previsão.

Ou seja, o módulo de *front-end* tem como principal responsabilidade a interação com o sistema de previsão, devendo também permitir a visualização e escolha de previsões, assim como a redação de documentos, o que implica funcionalidades de edição de texto. No protótipo a desenvolver será implementado um simples editor de texto que permitirá a interação com o sistema, no entanto, é importante deixar claro que o foco principal da presente dissertação é o *back-end*, ou seja, a previsão de texto.

Considerando a análise até agora realizada, é possível a identificação de algumas das principais operações que o sistema a desenvolver terá de suportar. Até o momento identificou-se a necessidade de criar uma base de conhecimento a partir de um conjunto de relatórios (treino); atualizar a base de conhecimento com um relatório finalizado (aprendizagem); e aceder a base de conhecimento para a realização de uma previsão (previsão).

Apesar de se tratarem de operações diferentes, internamente, o pré-processamento da informação em cada método será muito similar entre si, podendo esse processamento ser dividido em quatro passos distintos, sendo eles:

1. HTML *Decode* e *parse*;
2. Conversão para maiúsculas;
3. Classificação de *tokens*;
4. Remoção de espaços brancos extra;

Dado o facto de, atualmente, a maioria dos sistemas hospitalares serem aplicações *web*, os relatórios clínicos consistem na maioria das vezes em documentos HTML para possibilitar a sua visualização no *browser*. Há assim a necessidade de, ao processar os relatórios, realizar o *decoding* e *parsing* dos mesmos, de forma a remover as *tags* de HTML e extrair o título e corpo dos relatórios.

Seguidamente, deverá ser realizada uma conversão de todo o texto para maiúsculas. Este passo deve ser realizado de forma a reduzir o número total de possibilidades existentes na comparação de *strings*. Esta conversão apenas é possível pois a distinção entre maiúsculas e minúsculas não é importante no contexto do problema que a presente dissertação visa responder.

O terceiro passo consistirá na classificação de *tokens*, onde se atribuirá uma classe a determinadas palavras. Nos relatórios existem algumas informações que não são importantes para um sistema de previsão, por exemplo, uma data em si não tem utilidade nenhuma no sistema desenvolvido, no entanto, saber que uma data surge após aquele termo já pode ter alguma utilidade. Assim, sempre que uma data for identificada, a mesma deverá ser substituída por um *placeholder*, neste caso <*DATE*>. Proceder-se-á à identificação das seguintes classes de *tokens*:

- Datas;
- Volumes;
- Áreas;
- Comprimentos;
- Números;
- Nomes;

Para todos os casos, com exceção dos nomes, esta identificação será feita com recurso a expressões regulares. É importante realçar que os *tokens* que pertencem a uma classe poderão ter diversos

formatos, por exemplo, as datas poderão ser representadas por "*12/03*", "*12/03/2017*", "*12 de março*", entre outros formatos; devendo todas as expressões indicadas ser substituídas pelo respetivo *placeholder*. Relativamente a classificação dos nomes, a mesma será baseada na utilização de um simples dicionário.

Finalmente, o último passo consistirá na remoção dos espaços extra brancos, que também incluí *new lines*. Este processo permitirá a diminuição do tamanho dos dados e facilitará a divisão dos *tokens*.

De seguida é apresentada uma esquematização do pré-processamento descrito.

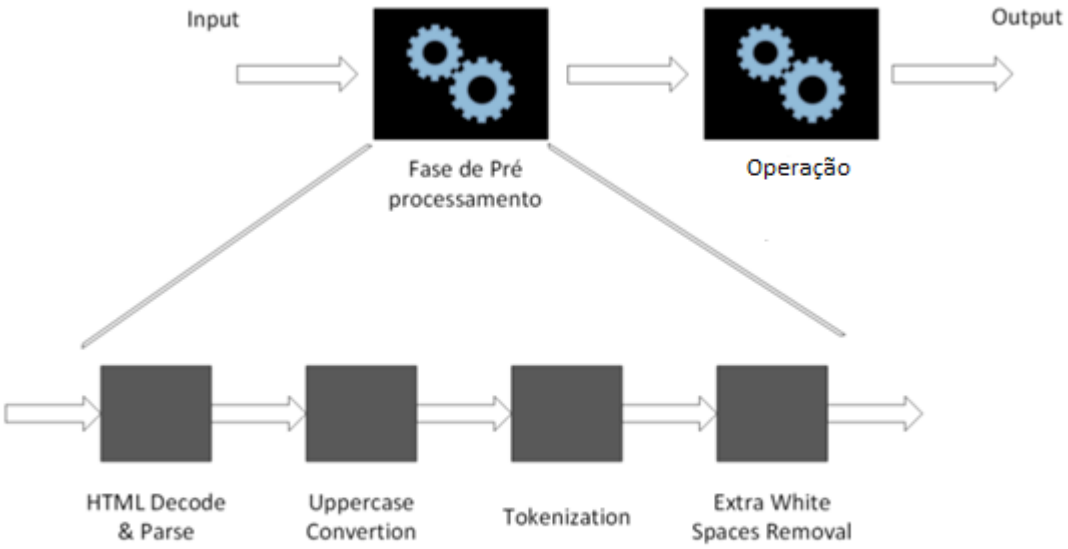

Figura 12 – Etapas do pré processamento.

Relativamente a base de conhecimento, a mesma será armazenada no sistema de ficheiros do *Windows*, no entanto, dependendo dos resultados do sistema, está planeada a sua passagem para uma base de dados.

Identificam-se assim todas as operações que o sistema a desenvolver pode realizar, sendo as mesmas enumeradas de seguida:

- Conversão de um conjunto de relatórios para uma base de conhecimento (treino);
- Atualização da base de conhecimento aquando a finalização de um relatório (aprendizagem);
- Previsão das próximas palavras tendo em conta o input do utilizador (previsão);
- Pré-processamento de relatórios;
- Operações de leitura e escrita de ficheiros;
- Operações de leitura e escrita da base de dados.

A partir da identificação das operações base do sistema, foi possível a criação de divisões lógicas de forma a organizar internamente a solução a desenvolver. A esquematização da arquitetura planeada é apresentada na figura 13.

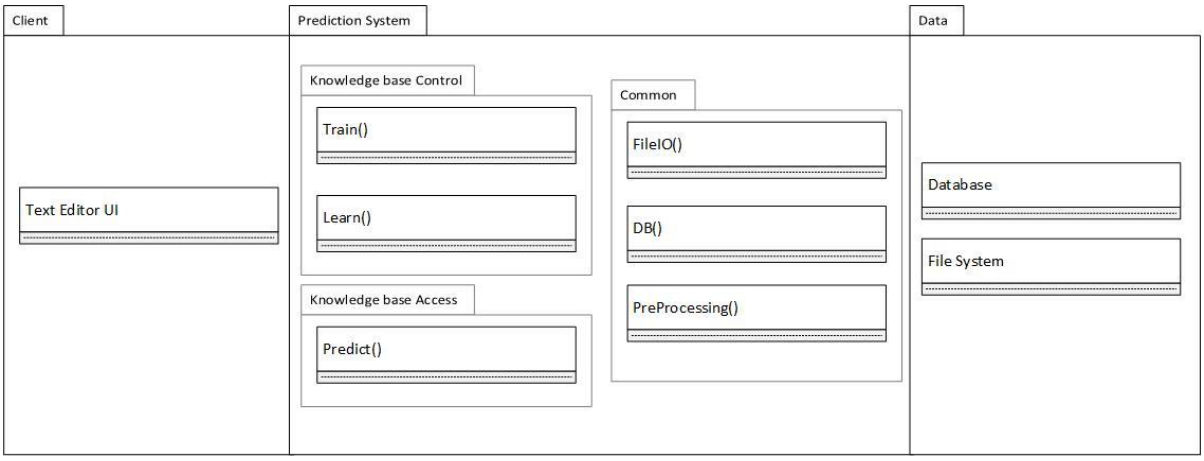

Figura 13 – Arquitetura do sistema.

Em suma, no que toca ao sistema de previsão, existem três módulos principais, sendo eles: o de controlo, que controlará toda a lógica que permita a alteração da base de conhecimento, aqui incluem-se os métodos de treino e aprendizagem; o de acesso, onde serão implementados métodos que necessitem de aceder (sem alterar) a base de conhecimento, como é o caso dos métodos de previsão; e o comum, que implementará todos os métodos comuns aos dois módulos anteriores, como por exemplo, funções de acesso à base de dados, a ficheiros e pré-processamento.

Após a identificação das operações base que permitem implementar o sistema planeado, e da esquematização da estrutura interna do sistema, considerou-se existirem elementos suficientes para iniciar o desenvolvimento da solução, sendo este processo apresentado no próximo capítulo, 5.

5 Desenvolvimento

Neste capítulo são apresentados os detalhes de desenvolvimento da solução anteriormente descrita. Inicialmente é apresentado o desenvolvimento em si, incluindo a descrição de alguns dos algoritmos utilizados. De seguida são apresentados alguns problemas de *performance* que se verificaram durante esta fase, assim como a respetiva resolução. Finalmente, é feita uma breve apresentação da configuração do sistema.

5.1 Desenvolvimento

O desenvolvimento do sistema de previsão decorreu dentro da normalidade, tendo-se seguido a arquitetura definida no subcapítulo 4.3. Os únicos problemas que se encontraram durante o processo de desenvolvimento estavam relacionados com a *performance* do sistema, tendo os mesmos sido resolvidos com recurso a algumas técnicas apresentadas no capítulo 5.2.

Relativamente à arquitetura do sistema em si, como anteriormente definido, foi decidido dividir o sistema em vários módulos, cada um com uma responsabilidade especifica. Um dos módulos criados foi o de controlo que, é responsável por todos os métodos que consigam alterar a base de conhecimento. Neste módulo estão incluídos todos os métodos relacionados com o treino e a aprendizagem.

Relativamente ao processo de treino, o mesmo deve ser executado apenas uma vez para a criação da base de conhecimento. Para tal, foi criada uma simples aplicação consola, onde se pode indicar: a fonte dos relatórios a processar (se da base de dados ou de um diretório); o profissional de saúde (caso a base de conhecimento a criar seja especifica para um utilizador); e o destino dos ficheiros que compõem a base de conhecimento. A *performance* e velocidade deste processo não são importantes para o desempenho da solução desenvolvida, isto devido ao facto de este passo ser apenas realizado uma única vez, antes da utilização do sistema, para a criação da base de conhecimento.

Como anteriormente referido, as bases de conhecimento criadas foram armazenadas no sistema de ficheiros do *Windows.* Relativamente a estrutura de dados utilizada, foi criada uma pasta para cada profissional de saúde onde, no seu interior, se encontram todos os dados necessários para as

previsões. Dentro dessa pasta, existem outras duas pastas que separam os dados utilizados para as previsões dos títulos e dos corpos. Depois, no último nível existem ainda mais duas pastas, uma com a listagem de *tokens*, e outra com a listagem dos *NGrams*. É de realçar que para a previsão apenas os conteúdos das pastas dos *NGrams* são utilizados, os conteúdos das pastas *tokens* consistem apenas em informação estatística utilizada para *debug*. Resumindo, a estrutura utilizada para o armazenamento dos dados pode ser representada pela seguinte *file tree*:

```
> ID do profissional de saúde   (pasta)
.   > títulos                   (pasta)
.   .       > tokens            (pasta)
.   .           …               (ficheiros – apenas para debug)
.   .       > ngrams            (pasta)
.   .           …               (ficheiros)
.   > corpos                    (pasta)
.   .       > tokens            (pasta)
.   .           …               (ficheiros – apenas para debug)
.   .       > ngrams            (pasta)
.   .           …               (ficheiros)
```

No subcapítulo 5.2.2 será explicado com maior detalhe os ficheiros no interior da pasta "*ngrams*", sendo que, neste momento, será suficiente saber que o mesmo consiste em inúmeros ficheiros de *NGrams*, onde cada ficheiro armazena uma lista de *NGrams*, assim como a respetiva frequência. Esta frequência é utilizada no processo de previsão, aquando o calculo da probabilidade do *NGram*.

Voltando ao processo de treino, o mesmo pode ser dividido nos seguintes quatro passos:

1. Leitura dos relatórios;
2. Pré-processamento dos relatórios;
3. Criação dos *NGrams*;
4. Criação da estrutura de dados;

O primeiro passo realizado aquando o processo de treino é a leitura dos relatórios disponíveis. Seguidamente, a cada documento é-lhe aplicado o pré-processamento, que consiste nas seguintes fases: HTML *decode* e *parse*; conversão para maiúsculas; classificação de *tokens*; e remoção de espaços brancos extra; este processo foi apresentado com maior detalhe no capítulo 4.3.2. O objetivo do pré-processamento é a uniformização e preparação da informação para o passo seguinte, que consiste na criação dos *NGrams*. Por exemplo, se o sistema estiver configurado para durante o processo de treino gerar *NGrams* de ordem 4, se a frase, "Verificam-se evidências de nódulos no exame de 12/03/2017" for fornecida ao sistema, serão criados os seguintes *NGrams*:

- VERIFICAM-SE EVIDÊNCIAS DE NÓDULOS
- EVIDÊNCIAS DE NÓDULOS NO
- DE NÓDULOS NO EXAME
- NÓDULOS NO EXAME DE
- NO EXAME DE *<DATE>*

Finalmente, após a geração de todos os *NGrams*, os mesmos são adicionados aos respetivos ficheiros de *NGrams*. Se estes não existirem o respetivo ficheiro será criado, caso contrário, a sua frequência é atualizada. Após todo este processo a base de conhecimento foi criada, estando pronta a ser utilizada.

Relativamente ao processo de aprendizagem, o mesmo implica a atualização da base de conhecimento tendo em conta um relatório finalizado pelo utilizador. Também aqui não se prestou atenção ao desempenho do processo, pois a execução desta operação não é realizada durante a utilização do sistema de previsão. O processo de aprendizagem é em tudo muito similar ao de treino, consistindo o mesmo nos seguintes passos:

1. Pré-processamento do relatório;
2. Criação dos *NGrams*;
3. Atualização da estrutura de dados.

Inicialmente o relatório finalizado é sujeito à fase de pré-processamento, isto para preparar os dados para a criação dos *NGrams*, sendo estes posteriormente adicionados a base de conhecimento.

O outro módulo identificado durante o planeamento do sistema é o de acesso que, é responsável por todos os métodos que acedem (sem alterar) a base de conhecimento, como é o caso dos métodos relacionados com a previsão de texto. Aqui, ao contrário do que aconteceu até agora, deseja-se que este processo seja o mais rápido possível, tendo-se para tal, utilizado algumas estratégias que serão apresentadas no subcapítulo 5.2.2. Relativamente ao processo em si, o mesmo pode ser dividido nas seguintes fases:

1. Pré-processamento do relatório;
2. Obtenção do contexto;
3. Obtenção dos *NGrams*;
4. Obtenção das previsões.

O algoritmo utilizado para a realização da previsão de texto recebe como argumento todo o texto introduzido pelo utilizador no editor de texto. O primeiro passo realizado é o pré-processamento, fase já abordada anteriormente. Após este processo são selecionados os últimos *tokens* do texto inserido, ou seja, por outras palavras, é obtido o contexto da frase a ser redigida. Com o contexto extraído, é possível verificar se o utilizador terminou de escrever a última palavra. Este passo é importante pois caso não tenha terminado existe um número maior de possibilidades para a previsão. Este número de possibilidades esta relacionada com a forma como a base de conhecimento está estruturada. De forma a facilitar as pesquisas de *NGrams*, foi definido como identificador do *NGram* o primeiro *token*, esse identificador dará origem a um ficheiro com todos os *NGrams* que tenham esse identificador. Por exemplo, considerando os seguintes *NGrams* de ordem 3, “verificam evidências de” e “verifica evidências de”, os dois serão guardados em ficheiros diferentes, já que os *tokens* utilizados para a identificação de ambos são diferentes. No caso de o utilizador não ter completado a última palavra e ter escrito “verifica”, os *NGrams* dos dois ficheiros serão utilizados para a previsão, no entanto, se o utilizador terminou a palavra, apenas o ficheiro do segundo *NGram* será utilizado. Assim, tendo em conta a última palavra do utilizador, são pesquisados todos os *NGrams* que possam corresponder a esse *token*. No caso do número mínimo de *NGrams* não ter sido atingido, o contexto da palavra é

reduzido e a pesquisa volta a ser realizada, sendo os novos resultados adicionados aos anteriores. Este processo é repetido até que o número mínimo de *NGrams* seja atingido, ou não seja possível reduzir mais o tamanho do contexto. De seguida os *NGrams* obtidos são ordenados por ordem de preferência tendo em conta a probabilidade e o tamanho do contexto. É necessário compreender que nesta fase, parte dos *NGrams* selecionados consistem no contexto utilizado para a sua pesquisa, no entanto, a parte correspondente ao contexto já se encontra no relatório, não sendo a mesma útil para a apresentação de sugestões ao utilizador. De forma a resolver este problema, o contexto é retirado dos *NGrams* selecionados, sendo o resultado um conjunto de *NGrams* parciais. Finalmente, volta-se a realizar a pesquisa previamente referida, mas agora com estes *NGrams* parciais, obtendo-se a continuação da frase que se encontra a ser redigida pelo utilizador. Dos *NGrams* obtidos são depois selecionadas as palavras e frases a serem apresentadas como sugestões. No caso de qualquer alteração do *input* do utilizador, este processo volta a ser calculado de forma a oferecer novas sugestões.

Na figura 14 é apresentado um diagrama de atividades relativo ao processo acima descrito.

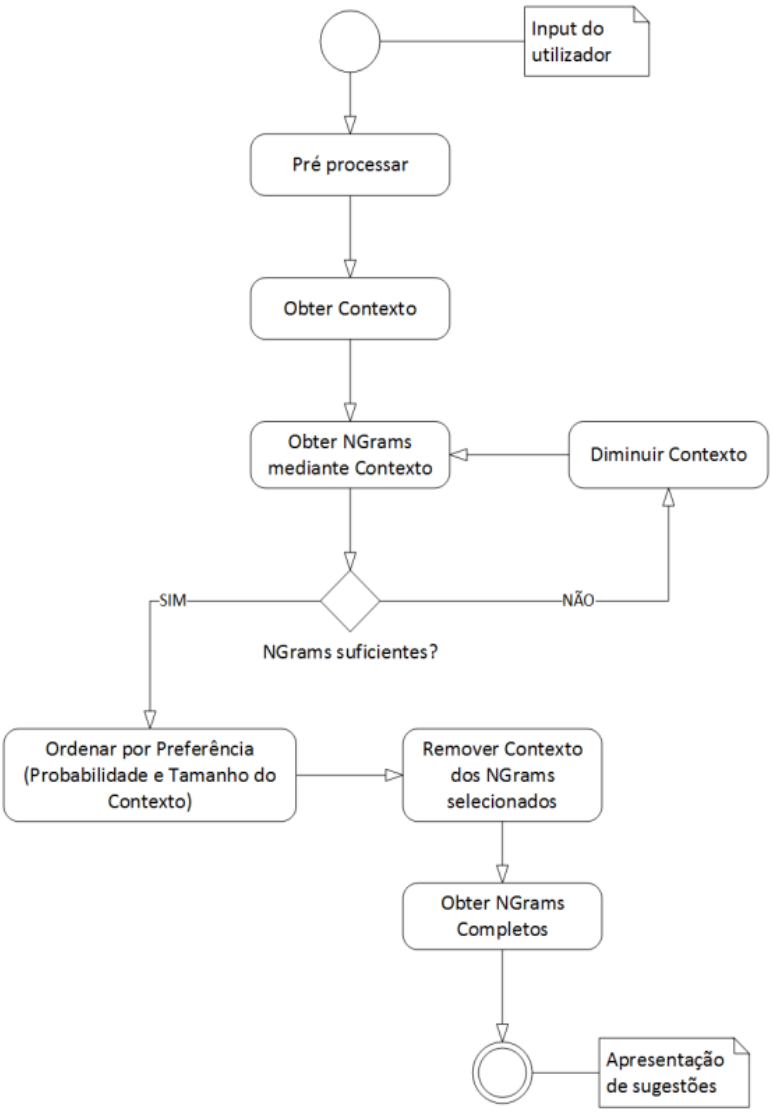

Figura 14 – Diagrama de atividades relativo ao processo de previsão.

Finalmente, o último módulo criado é o *comum*, que como o nome indica contem todos os métodos comuns aos dois módulos apresentados anteriormente. Este módulo incorpora: os métodos relacionados com o pré-processamento; os métodos que realizam operações sobre ficheiros; e os métodos de acesso à base de dados.

5.2 *Performance*

Como anteriormente referido, durante o processo de desenvolvimento da solução proposta foram encontrados dois problemas que punham em causa a *performance* do sistema. Os problemas encontrados podem ser divididos em duas categorias: problemas relacionados com memória; e problemas relacionados com a lentidão geral do sistema. Nos próximos subcapítulos iremos explicar como foram resolvidos ambos os problemas.

5.2.1 Memória

O problema de memória que se encontrou durante o desenvolvimento do sistema estava relacionado com o consumo excessivo da RAM. Felizmente, o problema foi rapidamente detetado, assim como a sua origem. No inicio do desenvolvimento da solução, todos os relatórios processados eram carregados para a memória, mas nunca eram removidos pelo *garbage collector* (GC), o que dava origem a um elevado consumo de memória. Como alternativa, passou-se a carregar os relatórios por *batchs*, ou seja, passou-se a carregar uma pequena quantidade de relatórios de cada vez. Quando o processamento dos mesmos terminava, os mesmos eram removidos da memória, carregando-se de seguida um novo grupo de relatórios. Desta forma, conseguiu-se manter o consumo de memória abaixo dum nível aceitável.

5.2.2 Processamento

Os problemas relacionados com a lentidão do sistema foram mais difíceis de resolver. A resolução deste problema era obrigatória, visto afetar seriamente os tempos de previsão do sistema. Para tal, foram utilizadas quatro técnicas que melhoraram a performance do sistema.

Uma das soluções utilizadas para tentar acelerar este processo, e já anteriormente apresentada, foi a criação de identificadores para os *NGrams*, guardando todos os *NGrams* com o mesmo identificador no mesmo ficheiro. Em parte, este processo ajudou na redução do tempo de previsão acelerando a pesquisa dos *NGrams*. Foi também utilizado processamento paralelo nas pesquisas dos *NGrams*. Tal é possível visto que cada pesquisa é independente entre si, não havendo criação de conflitos (*race conditions* e *deadlocks*). Outra solução também desenvolvida foi a criação de uma cache para guardar os *handlers* dos ficheiros mais utilizados, de forma a não estar constantemente a abrir o mesmo ficheiro, processo que pode gastar tempo precioso quando se deseja ser o mais rápido possível. Finalmente, a última técnica utilizada, e que realmente fez a diferença no tempo de previsão, foi a utilização de dicionários com *hash tables*. Graças a utilização de *hashes* nos dicionários todas as operações de pesquisa, inserção, remoção, entre outras, passaram a ter uma complexidade O(1)

(Microsoft, 2017). As utilizações destas técnicas permitiram aumentar a velocidade do sistema ao ponto de o mesmo poder ser utilizável e testável.

5.3 Configuração do sistema

O sistema desenvolvido pode ser configurado através da edição do ficheiro de configuração da aplicação. As principais configurações para o sistema de previsão são:

- Tamanho do contexto: corresponde ao número de *tokens* utilizado no contexto do método de previsão;
- Tamanho dos *NGrams*: corresponde ao número de *tokens* utilizado nos *NGrams* no método de previsão;
- *Encoding*: *Encoding* utilizado em ficheiros;
- Diretório dos *NGrams*: Caminho onde os ficheiros da base de conhecimento estão guardados;

6 Avaliação

Neste capítulo será apresentado todo o processo de avaliação do sistema desenvolvido. Inicialmente serão apresentados os aspetos que foram avaliados. Seguidamente abordar-se-ão as hipóteses testadas, e as metodologias de avaliação, explicando os testes que foram realizados e a divisão dos dados para os mesmos. Finalmente, são apresentados os resultados obtidos e respetiva análise.

6.1 Métricas

Resumindo sucintamente o trabalho realizado, o mesmo consiste num protótipo que permite a previsão de texto mediante as palavras redigidas pelo utilizador. Esse protótipo será dividido em dois módulos: o *back-end*, que será responsável por todos os métodos relacionados com previsão, treino e aprendizagem; e a *front-end*, que será a *interface* entre o utilizador e o sistema em si. É importante realçar que o principal âmbito da presente dissertação é a previsão de texto, sendo por isso o principal foco da avaliação o *back-end*, no entanto, será também apresentada uma possível forma de avaliar a *front-end*, ainda que, na prática, não se tenha procedido à avaliação da mesma. Cada um destes módulos tem objetivos diferentes. O módulo de *back-end* deve ter como principal objetivo a correta previsão dos próximos termos a serem utilizados, de preferência o mais rápido possível, e tendo em conta a aprendizagem do sistema. Quanto à *front-end*, esta deverá providenciar ao utilizador uma forma prática, rápida e eficaz de escolher sugestões e redigir o relatório. Assim, face às diferentes responsabilidades de cada módulo, optou-se pela utilização de um conjunto de métricas diferentes para a sua avaliação.

Desta forma, para a avaliação da *front-end* poderia ser utilizada a seguinte métrica:

- Tempo de redação do relatório;

O tempo de redação de um relatório é uma métrica interessante para verificar a rapidez proporcionada pelo sistema de previsão, mas acima de tudo pela *interface* gráfica, no processo de redação de relatórios. Permite-nos por exemplo, identificar se a solução proposta trás alguma vantagem relativamente as soluções existentes no que toca ao tempo gasto durante a redação de

relatórios clínicos. Esta métrica pode ser representada pelo intervalo de tempo desde o inicio da redação de um relatório até ao momento em que este é finalizado.

Relativamente ao *back-end*, foram avaliados os seguintes indicadores de *performance*:

- Tempo de previsão;
- *Keystroke Savings* (KS);

Um aspeto importante de ser avaliado é o tempo útil de previsão do sistema. Por motivos óbvios, as sugestões devem ser apresentadas rapidamente de forma a não quebrar o raciocínio do utilizador. É possível obter o tempo útil de previsão tendo em conta o intervalo de tempo desde o último caractere introduzido pelo utilizador, até o momento da apresentação das sugestões. O segundo indicador apresentado é utilizado para verificar a eficácia das previsões efetuadas pelo sistema. Para tal, foi utilizado o conceito de *keystroke savings*, que é uma medida que permite saber a quantidade de *keystrokes* evitados graças à previsão de texto. Quanto mais elevado for este valor, mais eficaz e eficiente é o sistema no processo de previsão. A equação 10 mostra como se pode calcular a percentagem de KS (McCoy, 2008).

$$KS = \frac{Nº\ de\ keystrokes\ sem\ previsão - Nº\ de\ keystrokes\ com\ previsão}{Nº\ de\ keystrokes\ sem\ previsão} \times 100 \quad (10)$$

Com as métricas a avaliar identificadas, é agora possível a formulação das hipóteses a testar para a avaliação do sistema, hipóteses essas apresentadas no próximo subcapítulo.

6.2 Hipóteses

Após a identificação das métricas a utilizar para a avaliação do sistema, e tendo em conta os requisitos definidos no capítulo 4.3.1, foi possível a definição das hipóteses a testar para cada uma das métricas definidas, hipóteses essas apresentadas de seguida:

1. O sistema deve apresentar uma taxa de *keystroke savings* superior a 75%;
2. O sistema deve demorar menos de 0.5 segundos para realizar uma previsão;
3. A utilização do sistema deve permitir reduzir o tempo de redação de relatórios em 25% comparativamente a outros métodos.

Das três hipóteses apresentadas, apenas as duas primeiras foram testadas, sendo apresentados os respetivos resultados no subcapítulo 6.4. O motivo de a terceira hipótese não ter sido testada, prende-se com o facto de que para validar essa hipótese, são necessários os tempos que os médicos demoram a redigir os relatórios nas diferentes abordagens identificadas no capítulo 1.3.1, informação essa que não se encontra disponível. A validação desta hipótese também implica que a *interface* gráfica esteja completamente desenvolvida, o que não se verificou dadas as limitações apresentadas no capítulo 7.

No próximo subcapítulo é apresentada a forma como as hipóteses formuladas foram testadas.

6.3 Metodologia de avaliação

Como referido anteriormente, das três hipóteses identificadas no capítulo anterior, apenas duas foram testadas. Para a realização destes testes foi desenvolvido um sistema automático que vai simulando a redação de um relatório clínico e que, utilizando o sistema de previsão desenvolvido, vai construindo o relatório tendo em conta as previsões realizadas. O sistema de testes desenvolvido vai registando o número total de *keystrokes* sem previsão, o número total de *keystrokes* com previsão, e ainda o tempo de cada previsão, sendo no final possível calcular a percentagem de *keystroke savings* e o tempo útil médio das previsões.

O sistema de testes desenvolvido deve, antes de realizar os testes, ser submetido a um processo de treino. Como é normal na avaliação de sistemas na área de IA que envolvam *machine learning*, os dados utilizados para a avaliação do sistema terão de ser diferentes dos utilizados durante o processo de treino. Se tal não se verificar, os resultados obtidos podem não ser considerados fidedignos já que, os mesmos foram obtidos tendo em conta situações que o sistema já processou. Tornasse assim necessário recorrer a técnicas de validação cruzada. Face à enorme quantidade de dados disponíveis, aproximadamente 120 000 relatórios clínicos, optou-se pela utilização do método *holdout.* Este método implica a divisão dos dados em dois grupos distintos, um para o treino, e outro para a avaliação em si. De todos os relatórios disponíveis foi selecionada uma amostra aleatória de 5 000 relatórios, todos eles provenientes do mesmo profissional de saúde. Essa amostra foi depois dividida em dois grupos de 2 500 relatórios cada, tendo-se utilizado o primeiro grupo para a realização do treino e o segundo grupo para a avaliação.

Outro fator que também se considerou durante o processo de avaliação foi a distinção entre os títulos e os corpos dos relatórios. Por norma, o titulo de um relatório é composto pelo nome do exame realizado e a zona do corpo que afeta. Isto faz com que a quantidade de títulos possíveis seja reduzida, pelo menos quando comparada com todas as possibilidades no corpo do relatório. Deste fator resulta que, naturalmente, a previsão de títulos seja mais eficaz e rápida que a previsão do corpo do relatório. Assim, de forma a evitar influenciar os resultados do sistema, optou-se pela separação destas duas secções do relatório na realização dos testes.

Finalmente, de forma a ser possível tirar conclusões sobre a relação entre os resultados obtidos e a variação de alguns dos parâmetros de configuração do sistema, foram definidas várias configurações que foram posteriormente testadas. Os parâmetros de configuração a que se deu particular atenção foram:

- O tamanho do contexto;
- O tamanho dos *NGrams*;
- E o número de sugestões.

Isto, tanto para a previsão dos títulos, como para a previsão dos corpos dos relatórios. É importante realçar que para que os resultados possam ser comparáveis, os dados utilizados para cada teste devem ser os mesmos, devendo a única variação destes testes residir nas configurações utilizadas. Os conjuntos de configurações, assim como os respetivos resultados obtidos são apresentados no próximo subcapítulo, 6.4, enquanto que a sua interpretação é apresentada no capítulo 6.5.

6.4 Resultados

Como anteriormente indicado, foram executadas várias séries de testes, com os mesmos dados, mas tendo em consideração parametrizações diferentes. Foram criadas quatro configurações base com determinados valores para o número de *tokens* utilizados no contexto e nos *NGrams* da previsão, e apresentando unicamente uma sugestão. De seguida, as mesmas configurações foram duplicadas, mas permitindo agora a apresentação de três sugestões. Um quadro resumo com as configurações que foram utilizadas nos testes é apresentado de seguida.

Tabela 7 – Quadro resumo das configurações testadas.

Identificação	Nº de *tokens* no contexto títulos	Nº de *tokens* no contexto corpos	Nº de *tokens* dos *NGrams* títulos	Nº de *tokens* dos *NGrams* corpos	Nº sugestões
A1	1	1	1	1	1
A2	1	1	1	1	3
B1	2	3	2	4	1
B2	2	3	2	4	3
C1	4	6	3	5	1
C2	4	6	3	5	3
D1	5	8	4	7	1
D2	5	8	4	7	3

Para cada conjunto de configurações definidas, foi simulada a elaboração de 2 500 relatórios, tendo sido registados os valores de KS (para os títulos, corpos e documentos completos) e ainda o tempo médio das previsões para cada teste. É importante também realçar que todos os testes realizados foram executados tendo em conta uma base de conhecimento gerada com *NGrams* de ordem 8. Os resultados obtidos são apresentados de seguida, sendo a sua análise realizada no subcapítulo 7.5.

Para a configuração A1 têm-se:

- Tamanho do contexto dos títulos: 1
- Tamanho do contexto dos corpos: 1
- Tamanho dos *NGrams* dos títulos: 1
- Tamanhos dos *NGrams* dos corpos: 1
- Número de sugestões: 1

Tendo-se obtido:

Tabela 8 – Resultados obtidos com a configuração A1.

Secção:	Nº de *keystrokes* sem previsão	Nº de *keystrokes* com previsão	KS (%)	Tempo de previsão (ms)		
				Médio	Máximo	Mínimo
Título	31 285	8 082	74.17	42	1 399	0
Corpo	731 476	224 005	69.38	139	5 192	0
Total	762 761	232 087	69.57			

Para a configuração A2:

- Tamanho do contexto dos títulos: 1
- Tamanho do contexto dos corpos: 1
- Tamanho dos *NGrams* dos títulos: 1
- Tamanhos dos *NGrams* dos corpos: 1
- Número de sugestões: 3

Tabela 9 – Resultados obtidos com a configuração A2.

Secção:	Nº de *keystrokes* sem previsão	Nº de *keystrokes* com previsão	KS (%)	Tempo de previsão (ms)		
				Médio	Máximo	Mínimo
Título	31 285	5 809	81.43	55	1 005	0
Corpo	731 476	160 551	78.05	139	5 875	0
Total	762 761	166 360	78.19			

Para a configuração B1:

- Tamanho do contexto dos títulos: 2
- Tamanho do contexto dos corpos: 3
- Tamanho dos *NGrams* dos títulos: 2
- Tamanhos dos *NGrams* dos corpos: 4
- Número de sugestões: 1

Tabela 10 – Resultados obtidos com a configuração B1.

Secção:	Nº de *keystrokes* sem previsão	Nº de *keystrokes* com previsão	KS (%)	Tempo de previsão (ms)		
				Médio	Máximo	Mínimo
Título	31 285	6 586	78.95	39	522	1
Corpo	731 476	86 844	88.13	271	166 130	0
Total	762 761	93 430	87.75			

Para a configuração B2:

- Tamanho do contexto dos títulos: 2
- Tamanho do contexto dos corpos: 3
- Tamanho dos *NGrams* dos títulos: 2
- Tamanhos dos *NGrams* dos corpos: 4
- Número de sugestões: 3

Tabela 11 – Resultados obtidos com a configuração B2.

Secção:	Nº de *keystrokes* sem previsão	Nº de *keystrokes* com previsão	KS (%)	Tempo de previsão (ms)		
				Médio	Máximo	Mínimo
Título	31 285	3 506	88.79	40	751	1
Corpo	731 476	79 644	89.11	221	182 558	0
Total	762 761	83 150	89.10			

Para a configuração C1:

- Tamanho do contexto dos títulos: 4
- Tamanho do contexto dos corpos: 6
- Tamanho dos *NGrams* dos títulos: 3
- Tamanhos dos *NGrams* dos corpos: 5
- Número de sugestões: 1

Tabela 12 – Resultados obtidos com a configuração C1.

Secção:	Nº de *keystrokes* sem previsão	Nº de *keystrokes* com previsão	KS (%)	Tempo de previsão (ms)		
				Médio	Máximo	Mínimo
Título	31 285	4 244	86.43	78	1 454	1
Corpo	731 476	86 195	88.22	721	373 282	0
Total	762 761	90 464	88.14			

Para a configuração C2:

- Tamanho do contexto dos títulos: 4
- Tamanho do contexto dos corpos: 6
- Tamanho dos *NGrams* dos títulos: 3
- Tamanhos dos *NGrams* dos corpos: 5
- Número de sugestões: 3

Tabela 13 – Resultados obtidos com a configuração C2.

Secção:	Nº de *keystrokes* sem previsão	Nº de *keystrokes* com previsão	KS (%)	Tempo de previsão (ms)		
				Médio	Máximo	Mínimo
Título	31 285	1 920	93.86	94	1 491	3
Corpo	731 476	77 656	89.38	983	426 364	1
Total	762 761	79 576	89.57			

Para a configuração D1:

- Tamanho do contexto dos títulos: 5
- Tamanho do contexto dos corpos: 8
- Tamanho dos *NGrams* dos títulos: 4
- Tamanhos dos *NGrams* dos corpos: 7
- Número de sugestões: 1

Tabela 14 – Resultados obtidos com a configuração D1.

Secção:	Nº de *keystrokes* sem previsão	Nº de *keystrokes* com previsão	KS (%)	Tempo de previsão (ms)		
				Médio	Máximo	Mínimo
Título	31 285	4 244	86.43	137	4 264	1
Corpo	731 476	86 078	88.23	2 226	214 436	0
Total	762 761	90 322	88.16			

Para a configuração D2:

- Tamanho do contexto dos títulos: 5
- Tamanho do contexto dos corpos: 8
- Tamanho dos *NGrams* dos títulos: 4
- Tamanhos dos *NGrams* dos corpos: 7
- Número de sugestões: 3

Tabela 15 – Resultados obtidos com a configuração D2.

Secção:	Nº de *keystrokes* sem previsão	Nº de *keystrokes* com previsão	KS (%)	Tempo de previsão (ms)		
				Médio	Máximo	Mínimo
Título	31 285	1 920	93.86	86	792	1
Corpo	731 476	77 823	89.36	1 154	404 558	3
Total	762 761	79 743	89.55			

6.5 Análise dos resultados

Tão importante como os resultados obtidos, é a sua correta análise. Para se proceder a essa análise é importante compreender a variação dos vários parâmetros nas diferentes configurações definidas. Foram especificadas quatro configurações base, A1, B1, C1 e D1, onde o tamanho dos contextos e dos *NGrams* utilizados aumenta gradualmente de A1 até D1. Estas configurações foram depois duplicadas, A2, B2, C2 e D2, onde a única diferença reside no número de previsões apresentadas. De seguida é apresentado um quadro resumo com os resultados obtidos para cada configuração.

Tabela 16 – Quadro resumo dos resultados obtidos.

Configuração	Secção	KS (%)	KS total (%)	Tempo de previsão médio (ms)*
A1	Título	74.17	69.57	42
	Corpo	69.38		139
A2	Título	81.43	78.19	55
	Corpo	78.05		139
B1	Título	78.95	87.75	39
	Corpo	88.13		271
B2	Título	88.79	89.10	40
	Corpo	89.11		221
C1	Título	86.43	88.14	78
	Corpo	88.22		721
C2	Título	93.86	89.57	94
	Corpo	89.38		983
D1	Título	86.43	88.16	137
	Corpo	88.23		2 226
D2	Título	93.86	89.55	86
	Corpo	89.36		1 154

*O tempo de previsão médio apresentado é sempre referente aos corpos. A razão de tal deve-se unicamente com o facto de serem os tempos mais longos.

Face as configurações e respetivos resultados apresentados, é interessante realizar dois tipos de análise com os dados obtidos. A primeira análise que pode ser interessante é uma comparação entre as configurações com os diferentes parâmetros, ou seja, a comparação de A1, B1, C1 e D1 com A2, B2, C2 e D2. A segunda passa pela análise da evolução dos resultados desde a configuração A até a configuração D. Para tal, e tendo em conta a tabela 16, foi possível a construção de dois gráficos que nos ajudarão na compreensão dos resultados obtidos. O gráfico 1 apresenta os valores de *keystroke savings* para os títulos, corpos e ainda o valor médio, das oito configurações definidas (A1, B1, C1, D1, A2, B2, C2 e D2). O gráfico 2 refere-se à evolução dos tempos médios de previsão para cada configuração. Os gráficos criados são apresentados de seguida.

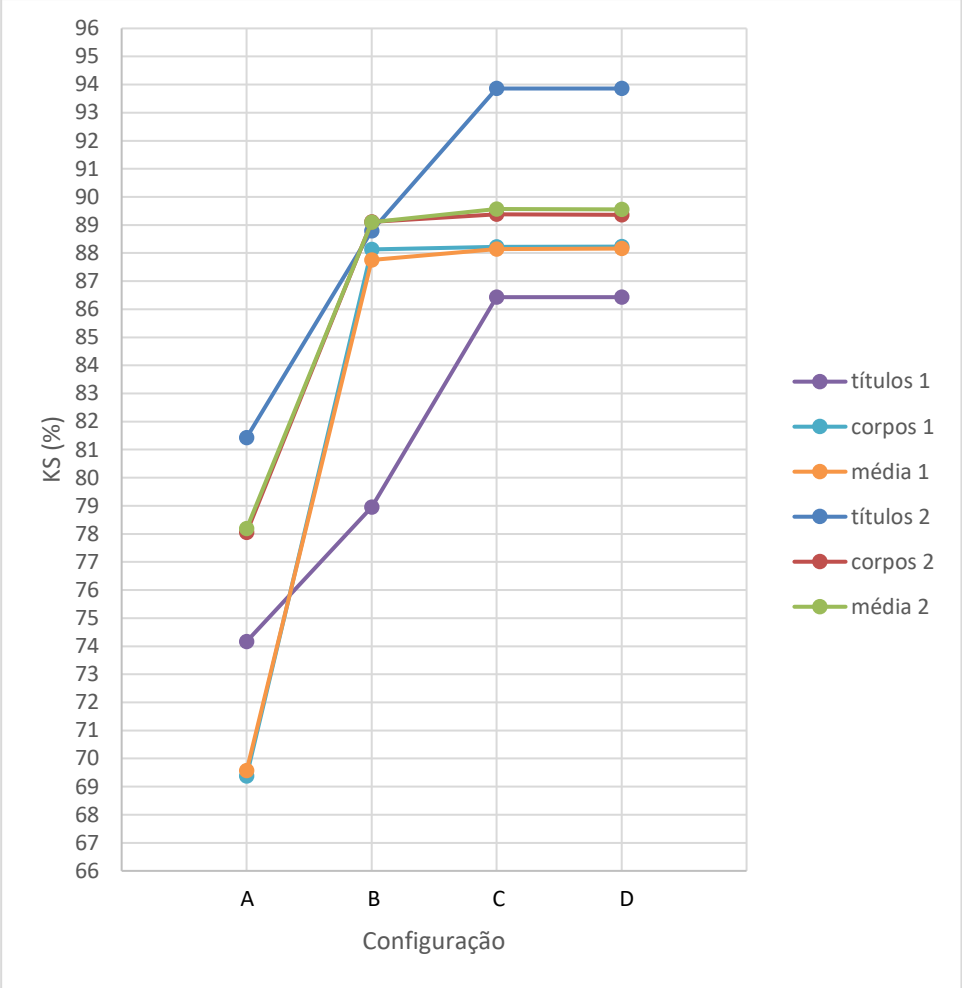

Gráfico 1 – KS dos títulos, corpos e documentos completos para todas as configurações.

No gráfico 1 são apresentados os valores dos *keystroke savings* para os títulos, corpos e documentos completos para cada configuração. As séries títulos 1, corpos 1 e média 1 correspondem respetivamente ao valor dos KS para os títulos, corpos e documentos completos das configurações A1, B1, C1 e D1, enquanto que as séries títulos 2, corpos 2 e média 2 correspondem as configurações A2, B2, C2 e D2. No gráfico 1 é possível verificar que à medida que o tamanho do contexto e *NGrams* utilizados na previsão aumenta, ou seja, passamos da configuração A para a D, também aumenta o valor dos *keystroke savings* obtidos. No entanto, é importante realçar que ao fim de algum tempo esse crescimento diminui, chegando mesmo a estabilizar e a ser nulo. Para o caso especifico dos títulos (séries títulos 1 e títulos 2) esse crescimento estabiliza na configuração C, enquanto que para os corpos dos relatórios (séries corpos 1 e corpos 2) o crescimento estabiliza na configuração B. No mesmo gráfico é também possível visualizar a diferença entre as configurações A1, B1, C1 e D1 com A2, B2, C2, e D2, ou seja, é possível compreender qual o efeito da variação do parâmetro que controla o número de sugestões apresentadas ao utilizador. Com o aumento do valor desse parâmetro, temos também um aumento do valor dos KS obtidos. Tal pode ser visualizado no gráfico 1 através da diferença entre as séries das configurações A1, B1, C1 e D1 (séries títulos 1, corpos 1 e média 1) e as configurações A2, B2, C2 e D2 (séries títulos 2, corpos 2 e média 2). Finalmente, no mesmo gráfico é ainda possível verificar que um pressuposto assumido anteriormente era falso. No subcapítulo 6.3 foi decidido separar os títulos e os corpos dos relatórios para avaliação da previsão. O motivo dessa

decisão está relacionado com o facto de se achar que a simplicidade da previsão dos títulos (pelo menos quando comparada com a dos corpos) poderia influenciar os resultados do sistema, pelo simples facto de, naturalmente, os mesmos apresentares melhores resultados que a previsão dos corpos dos relatórios. No entanto, no gráfico 1 é possível verificar que tal não é verdade, sendo claro que o valor da previsão do documento completo é muito similar, seguindo a mesma variação, do valor de previsão dos corpos (evolução dos pares de séries corpos 1/média 1 e corpos 2/média 2).

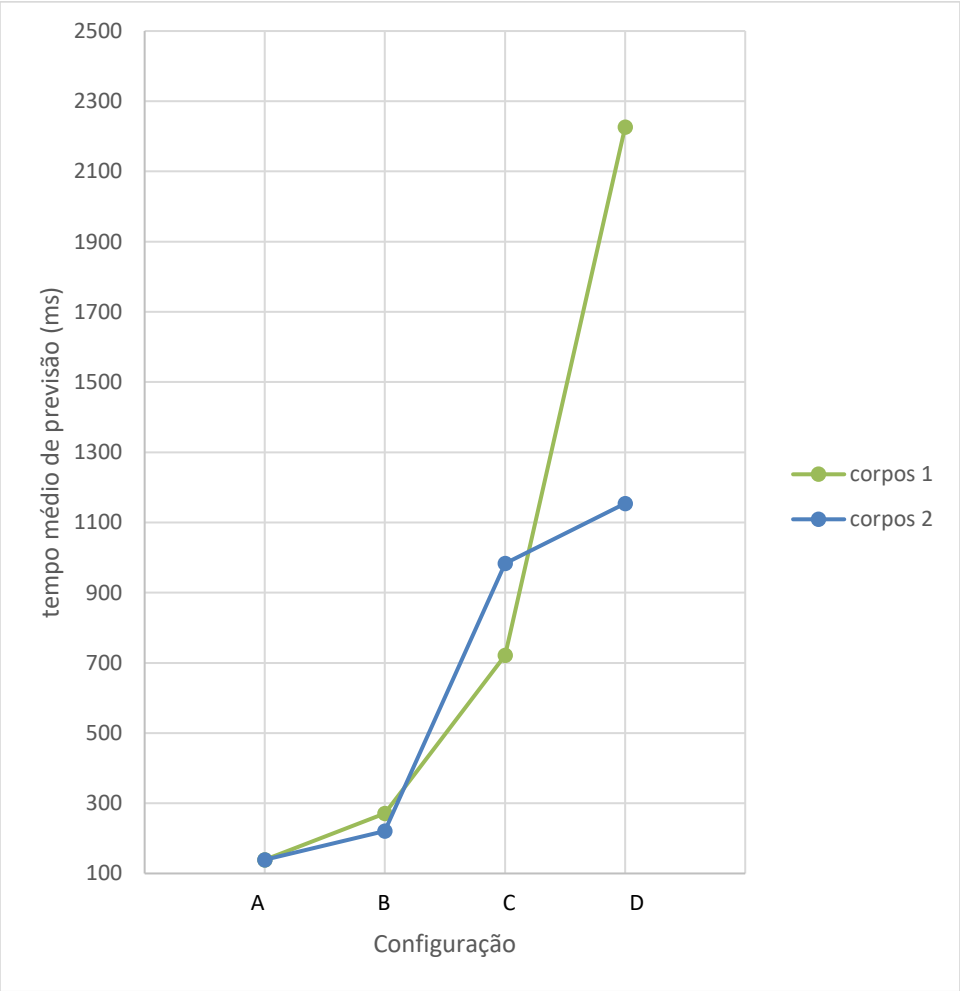

Gráfico 2 – Tempos médios (em milisegundos) das previsões para todas as configurações.

No gráfico 2 é possível visualizar a evolução do tempo médio das previsões dos corpos dos relatórios mediante a sua configuração. Os tempos médios dos títulos não são apresentados neste gráfico devido aos seus valores insignificantes quando comparados com os tempos dos corpos. A série corpos 1 representa os tempos médios de previsão obtidos pelas configurações A1, B1, C1 e D1, enquanto que a série corpos 2 corresponde as configurações A2, B2, C2 e D2. As principais conclusões a que se pode chegar com este gráfico é que o tempo médio de previsão aumenta, com o aumento do tamanho do contexto e *NGrams* utilizados na previsão, que corresponde à passagem da configuração A para a D. O outro ponto que é possível verificar é que o aumento do número de sugestões não implica um aumento do tempo médio de previsão. Tal pode ser visualizado através da diferença negativa entre as séries corpos 2 e corpos 1.

Face a análise efetuada, podemos concluir que para que o nosso sistema obtenha os melhores resultados possíveis, sem nunca sacrificar o tempo de previsão, podemos utilizar a configuração C2 para os títulos e a configuração B2 para os corpos. A configuração C2 para os títulos apresenta bons resultados, já estabilizados, e o seu tempo médio de previsão é ainda muito abaixo do limite estabelecido no subcapítulo 6.2. Relativamente aos corpos dos relatórios, apesar de a configuração B2 não apresentar os melhores resultados, estes são positivos, e o tempo médio é também ele abaixo do limite. A configuração C2 para os corpos apresenta algumas melhorias nos resultados dos KS, mas também implica um aumento considerável do tempo médio de previsão, não sendo por isso vantajosa a sua utilização. Com estas configurações (C2 para os títulos e B2 para os corpos), o sistema desenvolvido passa com uma margem confortável as hipóteses definidas no capítulo 6.2, podendo ser considerado um sistema eficaz e eficiente.

7 Conclusão

Neste capítulo são apresentadas as conclusões da presente dissertação. Inicialmente é apresentado um breve resumo abordando os pontos mais importantes dos capítulos anteriores. De seguida, são revistos os objetivos propostos, procedendo-se à análise da sua conclusão. Seguidamente são abordadas as limitações da solução desenvolvida, assim como em que consiste o trabalho a realizar no futuro. Finalmente, é realizada uma apreciação final, onde o autor dá a sua opinião pessoal sobre o trabalho desenvolvido.

7.1 Resumo

No inicio da presente dissertação é apresentado ao leitor um problema relacionado com os sistemas de informação hospitalares atualmente existentes. Este problema passa pelo consumo excessivo de recursos, tempo e dinheiro, comum a muitos hospitais e associado à fase de elaboração de relatórios. Este consumo excessivo de recursos é semelhante para qualquer uma das técnicas que se possa utilizar para a elaboração dos relatórios clínicos, independentemente de se utilizar uma técnica baseada em *templates* estáticos, ou uma técnica baseada na transcrição (manual ou automática) de áudio para texto. A partir do problema identificado, reconheceu-se imediatamente a existência de uma necessidade, tendo-se procurado uma solução que visa responder a mesma. O conceito desenvolvido consiste num sistema de previsão de texto, que suporte aprendizagem, e que possa ser integrado com um editor de texto. Durante a análise do estado da arte, verificou-se que os problemas relacionados com a previsão de texto pertencem a área de *language modeling*, existindo atualmente três abordagens diferentes para resolver problemas do género, sendo elas: abordagens baseadas em regras; abordagens baseadas em estatística e probabilidade; e abordagens hibridas. Durante a análise das três abordagens possíveis, assim como dos respetivos métodos disponíveis, destacaram-se as técnicas baseadas na utilização dos *NGrams* e das redes neuronais, especialmente pelos resultados positivos em trabalhos similares de outros autores. No final, com o recurso a uma ferramenta de tomada de decisão, AHP, foi decidido basear-se o trabalho desenvolvido no conceito de *NGrams*, principalmente pelo motivo anteriormente referido (resultados positivos) e pela relativa facilidade de implementação. De seguida, foi desenhado o sistema a implementar, procurando-se desenvolver uma solução modular, que fosse facilmente integrada noutros sistemas, e universal, para que não se

restringisse a uma cultura em específico. Durante o desenvolvimento da solução planeada, foram encontrados problemas relacionados com a *performance* do sistema, quer a nível de processamento, como a nível de memória. Estes problemas foram resolvidos evitando carregar toda a informação em memória, dividindo-a em unidades mais pequenas, e melhorando os métodos utilizados para pesquisas. Finalmente, relativamente a avaliação, procedeu-se a identificação das métricas e hipóteses a testar, tendo sido a previsão de texto (*back-end*) o principal foco dessa avaliação. Os resultados obtidos foram bastante positivos, tendo-se obtido taxas de *keystroke savings* superiores a 80%.

7.2 Objetivos realizados

No inicio da presente dissertação foram definidos e apresentados ao leitor uma série de objetivos que se pretendiam ver cumpridos. Ao longo deste documento, foi exposto todo o percurso que permitiu a resolução destes objetivos, sendo o grau de conclusão de cada um apresentado na seguinte tabela.

Tabela 17 – Grau de conclusão dos objetivos propostos.

Objetivo	Grau de conclusão
Aquisição e sintetização dos conhecimentos necessários	100%
Investigação do estado da arte de projetos similares	100%
Desenvolvimento de um modelo de previsão de palavras na elaboração de relatórios	100%
Desenvolvimento de um modelo de aprendizagem para incluir no modelo anterior	100%
Implementação dos modelos num protótipo	100%
Avaliação do protótipo	100%

Relativamente a avaliação do sistema, a mesma pode ser considerada positiva, tendo-se verificado as duas hipóteses formuladas no capítulo 6.3. Concluindo, o sistema desenvolvido pode ser considerado eficiente e eficaz, apresentando uma taxa de *keystroke savings* de aproximadamente 89%, e um tempo médio de previsão de 221 milisegundos.

7.3 Limitações e trabalho futuro

Apesar de todos os objetivos terem sido cumpridos com sucesso, e dos resultados obtidos terem sido positivos, a solução desenvolvida apresenta algumas limitações importantes de referir. Um dos maiores problemas da solução desenvolvida é a sua utilização por um profissional que seja novo na instituição de saúde e que nunca tenha redigido um relatório clínico. Os bons resultados do sistema verificados durante o processo de avaliação, partem do pressuposto que o sistema foi alvo de um processo de treino com alguns relatórios previamente redigidos por esse profissional de saúde. No

entanto, nas situações em que não existe uma base de relatórios prévia, esse processo de treino não pode ser realizado, dando origem a resultados negativos na previsão de texto. Nestas situações, a base de conhecimento é construída do nada, sendo naturalmente esperado, que no inicio da sua utilização, a eficiência e eficácia do sistema de previsão seja baixa. Como alternativa para este problema, poder-se-á criar uma base de conhecimento com os *NGrams* mais comuns entre diversos profissionais de saúde. Assim, os médicos que nunca tenham redigido um relatório clínico não terão de escrever a totalidade dos primeiros relatórios, existindo alguma informação na base de conhecimento que pode ser utilizada na previsão. Outra limitação do sistema desenvolvido esta relacionado com a sua dificuldade em lidar com erros ortográficos. Durante a elaboração de um relatório, é perfeitamente aceitável que um médico cometa erros ortográficos, por exemplo, escrevendo a palavra 'sao' quando queria escrever 'são'. Nestas situações, o sistema irá identificar duas palavras diferentes, que irão influenciar os resultados da previsão. Como agravante, existe ainda o facto de no final, os dados com os erros ortográficos serem utilizados no processo de aprendizagem, podendo as palavras com os erros serem utilizadas nos processos de previsão seguintes. Uma forma interessante de evitar este problema seria através da implementação de um conjunto de regras que permitiria tratar os caracteres 'á', 'à', 'ã' e 'â' como um 'a', ou seja, através da remoção dos acentos das palavras. No entanto, esta estratégia apenas resolveria os erros relacionados com palavras mal acentuadas. Uma forma mais correta de lidar com erros ortográficos seria através da utilização ou implementação de um corretor automático, tendo em atenção o contexto da palavra.

Relativamente ao trabalho a realizar no futuro, face os resultados positivos obtidos durante a avaliação do sistema desenvolvido, existe uma forte possibilidade de o mesmo vir a ser integrado em alguns produtos da FUJIFILM. Para além de alguns aspetos a melhorar indicados aquando a descrição das limitações do sistema, está atualmente a ser analisada uma possível integração da solução desenvolvida, com o atual sistema de reconhecimento de voz utilizado pela FUJIFILM no processo de elaboração de relatórios clínicos. Outra possibilidade que se encontra a ser estudada, é o possível desenvolvimento de um dispositivo que pode servir de teclado e rato ao mesmo tempo, permitindo ao utilizador escolher as previsões apresentadas pelo sistema ao mesmo tempo que utiliza o rato. Por norma, quando os médicos redigem um relatório relativo a um exame, eles utilizam uma das mãos para segurar no microfone, descrevendo o que visualizam nas imagens, enquanto que a outra mão controla o rato, onde vão realizando operações sobre as imagens, como por exemplo: passar para a imagem anterior/seguinte; *zoom*; criar anotações/formas; entre outras operações. O dispositivo a ser desenvolvido deve permitir ao utilizador realizar estas operações sobre as imagens, e ainda permitir de forma fácil escolher a sugestão da previsão desejada. No caso de o sistema desenvolvido vir realmente a ser utilizado, dado que o mesmo pode influenciar o diagnóstico de um paciente, poderá vir a ser necessário que o mesmo passe uma série de certificações, o que implica um percurso longo a validar a solução, com a criação de documentação oficial, e uma série de testes exaustivos.

7.4 Apreciação final

Todos os objetivos foram cumpridos com sucesso, assim como os resultados obtidos foram extremamente positivos. Tal pode ser devido a diversos fatores. Em primeiro lugar, os relatórios elaborados são redigidos num contexto muito especifico, hospitalar, existindo ainda a possibilidade de contextualizar os relatórios a um nível ainda mais detalhado, por serviços hospitalares (radiologia,

urologia, gastroenterologia, entre outros) e por autor. Este é provavelmente um dos principais fatores que influencia os resultados obtidos, e uma das principais vantagens quando comparado o sistema desenvolvido com outras soluções de previsão de texto (por exemplo, os teclados preditivos dos telemóveis). Outro fator que também é uma mais valia para a solução desenvolvida é a quantidade de relatórios que existe disponível, e que podem ser utilizados no treino do sistema.

Todo o trabalho realizado e apresentado ao longo da presente dissertação foi relevante para o desenvolvimento profissional do autor, assim como da empresa. Para o autor permitiu a exploração de uma área interessante, pouco ou nada abordada no ramo de Sistemas Gráficos e Multimédia do Mestrado de Engenharia Informática. Permitiu o alargar do conhecimento do autor, abrindo também novas possibilidades. Para a empresa, a presente dissertação permitiu abrir uma porta que permite a entrada da inteligência artificial no mundo dos sistemas de informação hospitalares, mais particularmente na área de *reporting*.

No geral, todo o trabalho desenvolvido foi positivo, com ótimos resultados, sendo certamente uma mais valia para todos.

Referências

Abascal, J., & Vitoria, N. (1994). Using statistical and syntactic information in word prediction for input speed enhancement. *Information Systems Design and Hypermedia* (pp. 223-230). Anglet: Cépaduès Éditions. Obtido de https://www.researchgate.net/profile/Nestor_Garay-Vitoria/publication/258769528_Using_statistical_and_syntactic_information_in_word_prediction_for_input_speed_enhancement/links/0c960528f1331983df000000.pdf?origin=publication_detail

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2 de Março de 2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, pp. 1137–1155. Obtido de http://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf

Bickel, S., Haider, P., & Scheffer, T. (Outubro de 2005). Learning to complete sentences. *Proceedings of the European Conference on Machine Learning*, pp. 497-504. Obtido de https://www.cs.uni-potsdam.de/ml/publications/ecml2005-seco.pdf

Boretos, G. (Fevereiro de 2012). S-curves and their Applications in Marketing, Business, and the Economy. *MRA's Alert!* Obtido de http://www.forecastingnet.com/Alert_0212_34-39.pdf

Cambridge University. (7 de Abril de 2009). *Tokenization.* Obtido de Cambridge University - Natural Language Processing: http://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html

Chiticariu, L., Li, Y., & Reiss, F. (Outubro de 2013). Rule-based Information Extraction is Dead! Long Live Rule-based Information Extraction Systems! *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 827-832). Seatle: Association for Computational Linguistics. Obtido de https://aclweb.org/anthology/D/D13/D13-1079.pdf

Choplin, R., Boehme, J., & Maynard, C. (Janeiro de 1992). Picture Archiving and Communication Systems: An Overview. *RadioGraphics, 12*.

Chowdhury, G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, pp. 51-89.

Conselho Europeu das Ordens dos Médicos (CEOM). (10 de Junho de 2011). *Carta Europeia de Ética Médica.* Obtido em 28 de Janeiro de 2017, de Ordem dos Médicos: https://www.ordemdosmedicos.pt/?lop=conteudo&op=217eedd1ba8c592db97d0dbe54c7adfc&id=9719a00ed0c5709d80dfef33795dcef3

Copestake, A. (2004). Corpora. Em A. Copestake, *Natural Language Processing* (p. 19). Obtido de https://www.cl.cam.ac.uk/teaching/2002/NatLangProc/revised.pdf

DICOM. (s.d.). *What is DICOM?* Obtido de DICOM: http://dicom.nema.org/dicom/geninfo/Brochure.pdf

Durão, C., Rocha, P., & Carvalho, N. (Setembro de 2003). Métodos de Entrada de Texto Especialmente Projectados para Comunicações Móveis. *Revista do DETUA, 4*, 20-31. Obtido de http://www.av.it.pt/nbcarvalho/docs/RevistaN7.pdf

Filipowicz, L. (26 de Setembro de 2016). *Mastering the Keyboard.* Obtido de iMore: http://www.imore.com/how-use-quicktype-keyboard-iphone-and-ipad

FUJIFILM. (2016). *Anual Report 2016.* Obtido em 1 de Fevereiro de 2017, de https://www.fujifilmholdings.com/en/investors/annual_reports/2016/pack/pdf_TOP/Annual-Report-2016.pdf

FUJIFILM. (2016). *CWM.* Obtido de SYNAPSE CWM: http://synapsecwm.com/contents.aspx?sid=25

FUJIFILM. (2017). *Company Profile*. Obtido em 1 de Fevereiro de 2017, de FUJIFILM: http://www.fujifilm.com/about/profile/

Gibaud, B. (31 de Agosto de 2009). The DICOM standard : a brief overview. 1-2. Obtido de https://hal.archives-ouvertes.fr/inserm-00344389/document

Henry, A. (15 de Junho de 2014). *Five Best Android Keyboards.* Obtido de LifeHacker: http://lifehacker.com/5922522/five-best-android-keyboards

HL7. (1987). *HL7 Mission.* Obtido de HL7: http://www.hl7.org/about/index.cfm?ref=common

Hutchins, J. (Novembro de 2005). The first public demonstration of machine translation: the Georgetown-IBM system, 7th January 1954. Obtido de http://www.hutchinsweb.me.uk/GU-IBM-2005.pdf

IBM. (8 de Janeiro de 1954). 701 Translator. *IBM Press release*. Obtido de https://www-03.ibm.com/ibm/history/exhibits/701/701_translator.html

Isokoski, P. (Abril de 2004). *Manual Text Input: Experiments, Models, and Systems.* University of Tampere, Department of Computer Sciences, Tampere. Obtido de http://tampub.uta.fi/bitstream/handle/10024/67373/951-44-5959-8.pdf;sequence=1

Jauregi, E., & Justel, D. (Agosto de 2007). Use of Tools, Methods and Techniques during the Fuzzy Front End of Inovation. *INTERNATIONAL CONFERENCE ON ENGINEERING DESIGN*, 3.

Jurafsky, D., & Martin, J. (1999). N-Grams. Em D. Jurafsky, & J. Martin, *Speech and Language Processing* (pp. 194-199). New Jersey: Prentice Hall,. Obtido de http://www.deepsky.com/~merovech/voynich/voynich_manchu_reference_materials/PDFs/jurafsky_martin.pdf

Kaminski, P., & Enachev, B. (2014). Obtido de Introdução ao Modelo de Negócio: CANVAS: http://sites.poli.usp.br/p/paulo.kaminski/INTRODU%C3%87%C3%83O%20AO%20BUSINESS%20MODEL%20CANVAS.pdf

Kannan, A., Kurach, K., Ravi, S., Kaufmann, T., Tomkins, A., Miklos, B., . . . Ramavajjala, L. (Agosto de 2016). Smart Reply: Automated Response Suggestion for Email. *KKD*. Obtido de

https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45189.pdf

Kay, M. (s.d.). *Suffix Trees.* Obtido de https://web.stanford.edu/~mjkay/suffix_tree.pdf

Koen, P. (25 de Maio de 2004). *Front End Innovation - Peter Koen.* Obtido de http://www.slideshare.net/BrandGenetics/koen-fei

Koen, P. (2014). *Front End Inovation.* Obtido de Front End Inovation: http://frontendinnovation.com/fei/what-is-the-new-concept-development-ncd-model

Koen, P. (2014). *What is the FEI?* Obtido de Front End Inovation: https://vimeo.com/94879899

Kriesel, D. (2005). Neural networks used for prediction. Em D. Kriesel, *A Brief Introduction to Neural Networks* (pp. 181-182).

MacKenzie, S. (2002). KSPC (Keystrokes per Character) as a Characteristic of Text Entry Techniques. *Proceedings of the Fourth International Symposium on Human Computer Interaction with Mobile Devices*, 195-210. Obtido de www.yorku.ca/mack/hcimobile02.html

Manning, C., & Schutze, H. (1999). The Ambiguity of Language: Why NLP Is Difficult. Em C. Manning, & H. Schutze, *Foundations of Statistical Natural Language Processing* (pp. 17-18). London: Massachusetts Institute of Technology.

Martins, M. (7 de Fevereiro de 2011). Processo Clínico Electrónico. *Faculdade de Engenharia da Universidade do Porto*. Obtido de https://repositorio-aberto.up.pt/bitstream/10216/58364/1/000146571.pdf

McCoy, K. (Junho de 2008). Evaluating Word Prediction: Framing Keystroke Savings. *Proceedings of ACL-08: HLT*, 261-264. Obtido de http://www.aclweb.org/anthology/P08-2066

Microsoft. (2017). *Hashtable.ContainsKey Method (Object) (System.Collections).* Obtido de Hashtable.ContainsKey Method (Object) (System.Collections): https://msdn.microsoft.com/en-us/library/system.collections.hashtable.containskey(v=vs.110).aspx

Nadkarni, P., Machado, L., & Chapman, W. (1 de Setembro de 2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association, 18*, pp. 544-551. doi:0.1136/amiajnl-2011-000464

Nandi, A., & Jagadish, H. (Setembro de 2007). Effective Phrase Prediction. *Proceedings of the 33rd international conference on Very large data bases*, pp. 219-230. Obtido de http://dbgroup.eecs.umich.edu/files/fcpaper07.pdf

Rosenthal, D., Bos, J., Sokolowski, R., Mayo, J., Quigley, K., Powell, R., & Teel, M. (Novembro de 1997). A Voice-enabled, Structured Medical Reporting System. *Journal of the American Medical Informatics Association*, 436-441.

Saaty, T. (2008). Decision making with the analytic hierarchy process. *Int. J. Services Sciences*, 86. Obtido de http://www.rafikulislam.com/uploads/resourses/197245512559a37aadea6d.pdf

Silva, R., & Belderrain, M. (26 de Agosto de 2005). *Considerações sobre Métodos de Decisão Multicritério.* Obtido de http://www.bibl.ita.br/xiencita/Artigos/Mec03.pdf

Swiffin, A., Arnott, J., Pickering, A., & Newell, A. (1987). Adaptive and predictive techniques in a communication prosthesis. *Augmentative and Alternative Communication , 3*, 181-191.

TouchChat. (s.d.). *VocabPC.* Obtido de TouchChat: http://touchchatapp.com/page-sets

Trinh, H., Waller, A., Vertanen, K., Kristensson, P., & Hanson, V. (Junho de 2012). Applying Prediction Techniques to Phoneme-based AAC Systems. *Proceedings of the Third Workshop on Speech and Language Processing for Assistive Technologies*, pp. 19-27. Obtido de http://delivery.acm.org/10.1145/2400000/2392859/p19-vertanen.pdf?ip=89.154.236.44&id=2392859&acc=OPEN&key=4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E6D218144511F3437&CFID=902016502&CFTOKEN=69929310&__acm__=1487453859_7428d6dc5c80a2a0153309f8

Turing, A. (1950). COMPUTING MACHINERY AND INTELLIGENCE. *Mind, 49*, 433-460. Obtido de https://www.csee.umbc.edu/courses/471/papers/turing.pdf

United Nations. (29 de Julho de 2015). *World population projected to reach 9.7 billion by 2050.* Obtido de United Nations - Department of Economical and Social Affairs: http://www.un.org/en/development/desa/news/population/2015-report.html

Vallez, M., & Jimenez, R. (2007). Natural Language Processing in Textual Information Retrieval and Related Topics. *Anuario Académico sobre Documentación Digital y Comunicación Interactiva*. Obtido de https://www.upf.edu/hipertextnet/en/numero-5/pln.html

Vitoria, N., & Abascal, J. (1997). A Syntatic Analysis-Based Word-Prediction Aid for People with Severe Motor and Speech Disability. *ACM* (pp. 241-244). Florida: ACM. Obtido de http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.225.9630&rep=rep1&type=pdf

Wang, A., Sable, J., & Spackman, K. (2002). The SNOMED Clinical Terms Development Process: Refinement and Analysis of Content. *AMIA*, 845-848. Obtido de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2244575/pdf/procamiasymp00001-0886.pdf

Weiss, S., & Indurkhyaa, N. (Dezembro de 1995). Rule-based Machine Learning Methods for Functional Prediction. *Journal of Artificial Inteligence*, pp. 383-403. Obtido de http://jair.org/media/199/live-199-1490-jair.pdf

Anexos

Anexo A – Cálculos intermédios AHP

Nesta secção são apresentados alguns cálculos intermédios efetuados durante a utilização do método AHP para decidir qual a abordagem seguir. Para cada um dos critérios é apresentada a tabela de comparação entre as várias abordagens, sendo de seguida apresentada a tabela normalizada, dividindo o elemento da tabela pela soma da respetiva coluna, assim como o respetivo vetor de prioridade relativa, que corresponde a média de cada linha.

Tabela 18 – Valores de comparação entre abordagens no critério complexidade.

Complexidade	Regras	*Ngrams*	RN	*Suffix Trees*
Regras	1.00	3.00	5.00	3.00
Ngrams	0.33	1.00	3.00	1.00
RN	0.20	0.33	1.00	0.33
Suffix Trees	0.33	1.00	3.00	1.00
Soma:	1.87	5.33	12.00	5.33

Tabela 19 – Valores de comparação do critério complexidade normalizados e prioridades.

Complexidade	Regras	*Ngrams*	RN	*Suffix Trees*	Prioridade relativa
Regras	0.54	0.56	0.42	0.56	0.52
Ngrams	0.18	0.19	0.25	0.19	0.20
RN	0.11	0.06	0.08	0.06	0.08
Suffix Trees	0.18	0.19	0.25	0.19	0.20

Tabela 20 – Valores de comparação entre abordagens no critério AA.

Suporta AA	Regras	*Ngrams*	RN	*Suffix Trees*
Regras	0	0	0	0
Ngrams	1.00	1.00	1.00	1.00
RN	1.00	1.00	1.00	1.00
Suffix Trees	1.00	1.00	1.00	1.00

Tabela 21 – Valores de comparação do critério AA normalizados e prioridades.

Suporta AA	Regras	*Ngrams*	RN	*Suffix Trees*	Prioridade relativa
Regras	0	0	0	0	0
Ngrams	0.33	0.33	0.33	0.33	0.33
RN	0.33	0.33	0.33	0.33	0.33
Suffix Trees	0.33	0.33	0.33	0.33	0.33

Tabela 22 – Valores de comparação entre abordagens no critério requer dados.

Requer dados	**Regras**	***Ngrams***	**RN**	***Suffix Trees***
Regras	0	0	0	0
Ngrams	1.00	1.00	1.00	1.00
RN	1.00	1.00	1.00	1.00
Suffix Trees	1.00	1.00	1.00	1.00

Tabela 23 – Valores de comparação do critério requer dados normalizados e prioridades.

Requer dados	**Regras**	***Ngrams***	**RN**	***Suffix Trees***	**Prioridade relative**
Regras	0	0	0	0	0
Ngrams	0.33	0.33	0.33	0.33	0.33
RN	0.33	0.33	0.33	0.33	0.33
Suffix Trees	0.33	0.33	0.33	0.33	0.33

Tabela 24 – Valores de comparação entre abordagens no critério requer dados.

Resultados	**Regras**	***Ngrams***	**RN**	***Suffix Trees***
Regras	1.00	0.33	0.25	0.50
Ngrams	3.00	1.00	0.50	2.00
RN	4.00	2.00	1.00	3.00
Suffix Trees	2.00	0.50	0.33	1.00

Tabela 25 – Valores de comparação do critério resultados normalizados e prioridades.

Resultados	Regras	*Ngrams*	RN	*Suffix Trees*	Prioridade relativa
Regras	0.10	0.09	0.12	0.08	0.10
Ngrams	0.30	0.26	0.24	0.31	0.28
RN	0.40	0.52	0.48	0.46	0.47
Suffix Trees	0.20	0.13	0.16	0.15	0.16