Universidade de Évora

Escola de Ciências e Tecnologia

Mestrado em Engenharia Informática

Deteção e Classificação de Sentimentos em Fontes de Informação não Estruturada

Hilário Barardo Fernandes

Orientador
José Miguel Gomes Saias

Évora, Março 2013

Mestrado em Engenharia Informática

Deteção e Classificação de Sentimentos em Fontes de Informação não Estruturada

Hilário Barardo Fernandes

Orientador
José Miguel Gomes Saias

Sumário

Uma das motivações que suporta a nossa constante busca por informação é a necessidade de saber "*o que os outros pensam?*". Com uma utilização cada vez maior das redes sociais, essa informação está finalmente ao alcance de empresas e outros organismos, permitindo que estas transformem esse conhecimento numa vantagem competitiva.

Nesta dissertação é apresentado um modelo para análise de sentimentos, que faz a deteção e classificação de referências a entidades em fontes de informação não estruturada, seguindo uma aproximação normalmente referida como de orientação semântica. Este recorre a um conjunto de regras cuja aplicação envolve a utilização de um léxico de sentimentos, e a exploração da relação de sinonímia entre palavras. Como resultado é indicado o tipo de sentimento (positivo, negativo ou neutro) e a entidade sobre a qual este incide.

Finalmente, é realizada uma avaliação do sistema desenvolvido num corpus em Português, procurando justificar as aproximações escolhidas e situar os resultados obtidos no panorama atual.

Sentiment Detection and Classification in Non Structured Information Sources

Abstract

One of the motivations that supports our constant search for information is the need to know "*what other people think?*". With the increasing use of social networks that information is finally within the reach of companies and other entities, allowing them to turn that knowledge into a competitive advantage.

In this dissertation we present a model for sentiment analysis, to detect and classify references to entities in non structured information sources, following a commonly named semantic orientation approach. The system uses a set of rules, which are applied with the aid of a sentiment lexicon, and explores the synonymity relationship between words. As a result, it identifies the type of sentiment present in the text (positive, negative or neutral) and the target to whom it applies.

Finally, we present an evaluation of the proposed system in a Portuguese corpus, looking to justify the approaches taken and compare the results obtained with the current state of the art systems.

Para a minha família e amigos.

Agradecimentos

A todas as pessoas que contribuíram de alguma forma para tornar este trabalho possível, deixo aqui uma dedicatória especial procurando expressar o meu agradecimento pelo seu contributo.

Em primeiro lugar gostaria de agradecer ao meu orientador, o Professor José Saias, pela orientação e motivação facultadas. A sua disponibilidade e experiência foram fundamentais durante as várias fases deste projeto, prontificando-se sempre para discutir as minhas escolhas e alertar-me das implicações associadas.

Um grande agradecimento ao meu colega e amigo Mário Mourão pela paciência para aturar as minhas birras, disponibilizando-se sempre para discutir problemas mesmo quando não fazia ideia do que eu estava a falar.

Para os meus colegas e amigos Ruben Silva, Miguel Reis e Artur Romão um grande obrigado. Sem todo o apoio, disponibilidade e motivação que sempre me deram este trabalho não teria certamente sido concluído. Um grande abraço para os três.

Um agradecimento especial à Cátia Barreto, por ter motivado o início deste projeto com a sua frontalidade e sinceridade características, simplesmente perguntando: "*Porque não acabas a tese?*". Por essa conversa e por todas as vezes que me pressionou para trabalhar desde então, o meu sincero obrigado.

Agradeço também à minha mãe e avó por sempre me apoiarem nas minhas decisões. Sei que posso sempre contar convosco quando precisar.

Não poderia deixar de mencionar o meu amigo Nuno Mesquita que me acompanhou na fase inicial deste projeto. A motivação obtida da nossa interação foi valiosa, como tal deixo aqui o meu agradecimento e um grande abraço.

Finalmente, agradeço a todos os meus colegas de trabalho e amigos não referidos anteriormente e que de alguma forma deram a sua contribuição. A sua ajuda e apoio foram o que tornou este projeto possível, para todos um grande abraço.

Acrónimos

IE Inteligência Empresarial

PLN Processamento de Linguagem Natural

EI Extração de Informação

AS Análise de Sentimentos

MO Mineração de Opinião

TA Tradução Automática

AA Aprendizagem Automática

TI Teoria da Informação

IA Inteligência Artificial

MUC Message Understanding Conference

REM Reconhecimento de Entidades Mencionadas

AM Análise Morfossintática

MaxEnt Maximum Entropy

SVM Support Vector Machines

RBEM Rule Based Emission Model

OS Orientação Semântica

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Nesta dissertação é descrito o trabalho efetuado numa área particular do Processamento de Linguagem Natural (PLN): a Análise de Sentimentos (AS). O principal objetivo é o desenvolvimento de uma ferramenta capaz de identificar sentimento em texto proveniente de fontes não estruturadas, relacionando este com o objeto sobre o qual incide. Ao longo dos capítulos deste documento são descritas várias metodologias aplicadas atualmente ao tema, identificada a aproximação escolhida e apresentados os resultados obtidos.

1.1 Enquadramento e Motivação

> *"What other people think" has always been an important piece of information for most of us during the decision-making process.*
>
> B. Pang and L. Lee [35]

Uma das motivações que suporta a nossa constante procura por informação é a tentativa de perceber "*o que os outros pensam?*" sobre algo ou alguma coisa. Com a crescente popularidade de recursos como os *sites* de opinião, os blogues pessoais e as redes sociais, surgem novas oportunidades de obter a resposta para essa pergunta. As pessoas agora podem e utilizam ativamente as tecnologias de informação para procurar e perceber a opinião de outros. A súbita explosão de atividade na área da Mineração de Opinião (MO) e AS, ocorre em parte como resposta ao crescente interesse em aproveitar todo um conjunto de informação disponível sem custos, mas de interpretação deveras complexa [35].

Muito antes da *World Wide Web* se tornar parte integrante do quotidiano de cada um, muitos de nós recorríamos a um amigo que nos recomendasse um mecânico ou nos dissesse

em quem pensava votar nas próximas eleições. Mas a *Internet* e a *web* tornaram possível descobrir opiniões e experiências de pessoas que nos são completamente desconhecidas, pessoas das quais nunca ouvimos falar e que colocam as suas opiniões *online* para serem consultadas por estranhos [35].

De acordo com duas pesquisas que envolveram 2000 Americanos adultos cada [14, 24]:

- Cerca de 81% dos utilizadores da *Internet* pesquisaram sobre um produto pelo menos 1 vez;
- 20% fazem-no no dia-a-dia;
- Daqueles que leram análises de restaurantes, hotéis e outros serviços, entre 73% e 87% dizem que este facto teve uma importância significativa na compra;
- Estes consumidores dizem estar dispostos a pagar entre 20% e 99% mais por um item com avaliação de 5 estrelas em vez de 4 (o valor depende do tipo de item considerado);
- 32% forneceram a sua opinião sobre um produto, serviço ou pessoa através de um sistema de classificação *online* e 30% colocaram um comentário *online* relativo a um produto ou serviço.

A grande procura dos utilizadores por conselhos e recomendações *online* que os dados anteriores revela, é apenas uma das razões para o aumento de interesse por sistemas que façam o tratamento de opiniões. No entanto Horrigan [36] chama a atenção de que mesmo com a maioria dos utilizadores americanos a reportarem experiências positivas, ao mesmo tempo 58% dizem não existir informação *online*, que foi impossível de encontrar ou que esta era confusa [35].

Do ponto de vista empresarial, a informação é nos dias que correm um bem precioso para quem se encontra numa posição de decisor. Uma decisão errada pode ter consequências devastadoras e como tal deve ser evitada a qualquer custo. Assim, a necessidade de ter acesso à informação que realmente importa e em tempo útil é fundamental para empresas e outros organismos cujo futuro depende das opções tomadas no presente. Dependendo do âmbito da empresa ou instituição em causa, a informação que importa obter poderá estar dispersa por diversos locais. Internamente, registos de vendas e faturação são normalmente considerados, e externamente, índices de bolsa ou estatísticas de acidentes são bons exemplos. Porém, atualmente isto não é suficiente, é necessário recolher informação dispersa em notícias, apresentações e páginas *web* sem formato definido. A Inteligência Empresarial (IE) requer que o máximo de informação relevante seja recolhida, tratada e relacionada, para que os factos e resultados significativos cheguem ao decisor. No entanto, muita dessa informação provém das fontes há pouco mencionadas, tornando assim a sua mineração e tratamento num problema para as ferramentas tradicionais de IE. Estas foram desenvolvidas e aperfeiçoadas ao longo do tempo para trabalhar com informação com formatos bem definidos, como tal é sem surpresa que se verificam resultados bastante diferentes quando

se deparam com grandes quantidades de informação sem estrutura fixa. Uma possível aproximação para alguns destes casos será aplicar técnicas de Extração de Informação (EI) e/ou PLN com o objetivo de chegar a uma versão estruturada da informação, esta poderá alimentar então os sistemas já existentes [30].

Nos objetivos de qualquer empresa deve sempre constar um cuja importância é indiscutível, a satisfação do cliente. Uma empresa que não trabalhe para melhorar a relação com os seus clientes certamente verá o impacto deste fator nos seus resultados. Mas como saber o que pensam os clientes? Existem alguns métodos atualmente em ampla utilização, questionários de satisfação e *user groups* sendo os mais comuns. Estes permitem a obtenção de informação relevante para a empresa de um modo estruturado, facilitando a sua utilização posterior. Os questionários podem ser direcionados e com um domínio bem definido para que as respostas se enquadrem dentro do previsto. Existem no entanto alguns problemas com estes métodos, a dificuldade associada à criação e manutenção dos questionários e no caso dos grupos de utilizadores, a promoção e logística associada para que sejam recolhidos resultados significativos possui também um custo considerável. Adicionalmente, a rigidez das perguntas impede que o cliente passe a sua opinião sobre um aspeto não abrangido pelo questionário ou o faça de um modo incompleto. Existe ainda o caso de utilizadores que não participam neste tipo de iniciativas, não permitindo assim recolher as suas opiniões [19].

O *feedback* espontâneo fornecido por atuais e possíveis clientes em blogues, via correio eletrónico, nas redes sociais, ou em *sites* que permitem a submissão de opinião contém a informação que as empresas tanto se esforçam por conseguir. No entanto esta não se encontra estruturada, encontra-se em texto livre sem quaisquer restrições o que no limite significa que pode nem existir uma opinião. Tendo em conta a quantidade disponível e o seu crescimento constante, os custos associados ao seu tratamento manual são exorbitantes, tornando a tarefa manual possível apenas com uma escolha cuidadosa das amostras a tratar [19].

Neste enquadramento é notória a necessidade de criar ferramentas capazes de extrair, estruturar e utilizar estas fontes de informação em crescimento. A investigação em áreas como o PLN, EI e finalmente a AS possui um papel fundamental na construção das bases necessárias para o aparecimento dessas ferramentas. Este trabalho surge neste âmbito procurando dar resposta a esta necessidade crescente e incontornável.

1.2 Objetivos e Contribuições

Esta dissertação procura definir uma solução capaz de responder às necessidades atuais de classificação de sentimento em texto. A solução desenhada deve ser capaz de identificar a presença de sentimento, o carácter desse sentimento e o alvo sobre o qual este incide. Como principais contribuições, este trabalho:

- Define uma arquitetura modular para um sistema extensível, capaz de classificar sentimento em texto escrito em linguagem natural.
- Propõe uma implementação desse sistema da qual fazem parte as seguintes técnicas:
  - Enriquecimento morfossintático prévio do texto.
  - Identificação de entidades mencionadas recorrendo ao reconhecimento de padrões, morfologia e utilização de um catálogo de menções.
  - Classificação de sentimento utilizando regras de sintaxe, léxicos de sentimento e morfologia.
- Implementa e avalia a solução proposta, identificando as suas vantagens e desvantagens, propondo direções futuras.

O estudo realizado incidiu principalmente sobre a análise de pequenos comentários, maioritariamente de foro político por conveniência, no entanto as técnicas empregues são válidas para outros domínios.

1.3 Organização da Dissertação

Esta dissertação estende-se ao longo de seis capítulos que devem ser abordados de forma sequencial e que se passam a descrever. O presente capítulo, Capítulo 1, contém uma introdução à temática abordada por este trabalho e a respetiva motivação para a sua realização. O Capítulo 2 descreve um conjunto de ferramentas e conceitos chave fundamentais para a compreensão das abordagens tidas na criação da solução proposta. O Capítulo 3 apresenta uma visão geral do estado atual dos sistemas de AS, procurando estabelecer o foco nas abordagens consideradas mais comuns. O Capítulo 4 apresenta a solução proposta, que foi alvo de implementação, e que resultou da investigação realizada. Os resultados obtidos pelo protótipo implementado são apresentados e discutidos no Capítulo 5. Finalmente, o Capítulo 6 apresenta as conclusões deste trabalho e futuras direções a considerar.

Capítulo 2

Conceitos e Ferramentas

2.1 Conceitos

A AS surge da aplicação do PLN, Linguística e EI para identificar e extrair a subjetividade presente em textos e outros recursos. Tendo em conta que o foco desta dissertação é precisamente a AS, torna-se importante para a sua compreensão conhecer alguns conceitos chave inerentes às metodologias e aproximações escolhidas.

2.1.1 Redes Sociais e *Microblogging*

O termo "rede social" é hoje vulgarmente utilizado para referenciar um tipo de serviço que tem atraído cada vez mais adeptos desde a sua aparição. Redes como o conhecido Facebook[1], o MySpace[2] ou o Google+[3], entre outros, foram sendo gradualmente incluídos no quotidiano do cidadão comum. Embora tecnologicamente os vários serviços sejam relativamente semelhantes, as comunidades que se juntam em torno de cada um podem ser diversificadas. Estas podem crescer baseadas nos interesses comuns entre indivíduos, na nacionalidade, crenças religiosas ou não estabelecer um foco em nenhum grupo em particular procurando cativar um público maior. Define-se o conceito de serviço de rede social como um sistema que permite realizar as seguintes ações [16].

1. Construir um perfil público ou parcialmente público dentro de um sistema fechado;
2. Gerir uma lista de outros utilizadores com quem partilha uma ligação;

[1] Rede social generalista em https://www.facebook.com.
[2] Rede social vocacionada para o entretenimento. Acessível em http://www.myspace.com.
[3] Rede social da Google. Acessível em https://plus.google.com.

3. Visualizar e navegar a sua lista de ligações e as de outros dentro do sistema.

Paralelamente ao crescimento dos serviços de rede social, um outro tipo de sistema emergiu garantindo a sua posição e a sua própria comunidade de utilizadores neste mercado, o *microblogging*. Este tipo de serviço permite aos utilizadores publicar pequenas mensagens (normalmente até 200 caracteres) descrevendo o seu estado atual. A simplicidade do conceito permite que estas sejam partilhadas com amigos ou interessados através de mensagens instantâneas, telemóveis, *email* ou através da Internet. Um dos mais populares serviços de *microblogging* é o Twitter[4], permitido seguir as *tweets*[5] de alguém e partilhar as nossas com quem nos segue [25].

2.1.2 Processamento de Linguagem Natural

O PLN tem como principal objetivo o estudo computacional da linguagem humana, e através deste, analisar e desenhar agentes computacionais que a utilizam para adquirir informação de outros agentes, humanos ou não, permitindo-lhes assim a interagir e alterar o estado do mundo [21].

Desde o aparecimento da Inteligência Artificial (IA) que um dos seus principais objetivos foi o desenvolvimento de métodos para compreender a língua natural [8]. Nas décadas de 50 e 60 a investigação incidiu principalmente sobre a Tradução Automática (TA). Os primeiros resultados transpareceram imediatamente a dificuldade da tarefa, mesmo em pequenos exemplos uma palavra poderia ser traduzida incorretamente se o contexto em que ocorre não fosse considerado [23]. Chegou-se assim rapidamente à conclusão de que a compreensão da linguagem iria requerer mais do que informação lexical e gramatical. Seria necessário ter informação semântica e possuir um conhecimento geral do mundo.

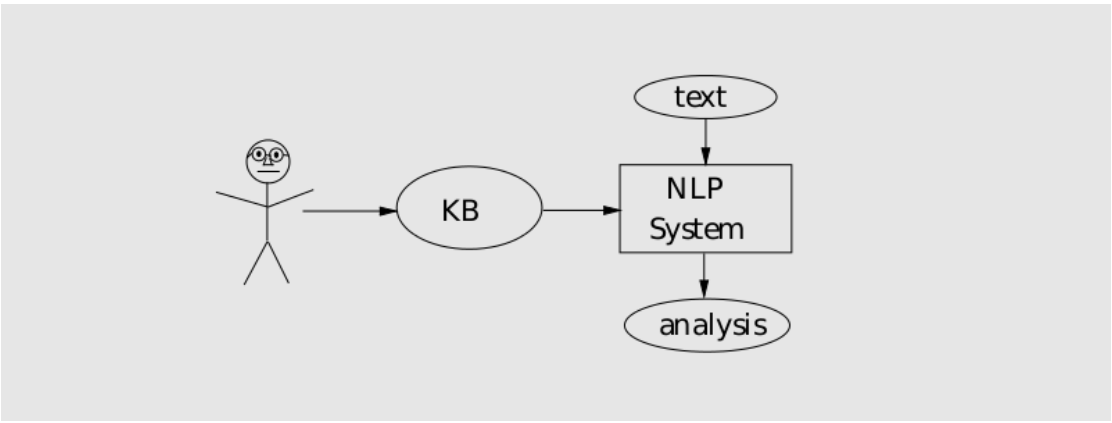

Figura 2.1: Método racional ou lógico. [8]

[4]Serviço de *microblogging* acessível em https://twitter.com.
[5]Nome dado a uma mensagem enviada através do Twitter.

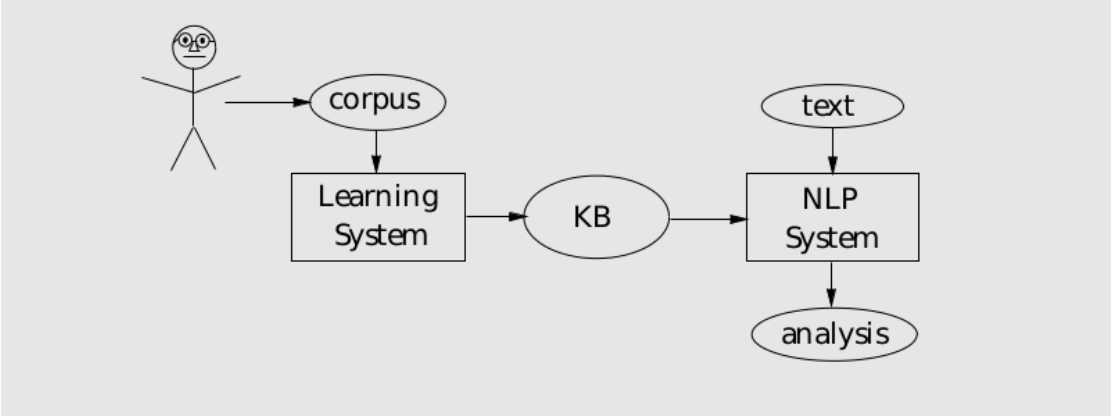

Figura 2.2: Método empírico ou prático. [8]

Nos anos 70 a vertente que mereceu maior atenção foi a IA, a investigação focou-se principalmente no conhecimento sobre o mundo e o seu papel na construção e manipulação de representações de significado. Mesmo com os problemas que persistiram da década anterior surgiram sistemas de IA que demonstraram aspetos interessantes de compreensão da linguagem em domínios restritos [52, 53, 50]. Na década de 80 era já consensual o facto de que construir sistemas de PLN fiáveis e extensíveis era uma tarefa muito mais complicada do que o inicialmente previsto, mesmo para aplicações com um domínio bastante restrito. Como consequência dos resultados poucos satisfatórios obtidos por sistemas que assentavam numa abordagem empírica (figura 2.2), e respondendo ao estímulo fornecido pelo desenvolvimento da teoria gramatical na Linguística, deu-se um progresso contínuo na criação de sistemas de língua natural recorrendo a gramáticas simbólicas criadas manualmente e a bases de conhecimento lógicas (figura 2.1) [3]. No entanto, o desenvolvimento destes sistemas continuava a ser difícil, requerendo um enorme conhecimento dos domínios a que se aplicavam. Adicionalmente, eram propícios a falhas e não funcionavam adequadamente fora das tarefas específicas para que eram desenhados. Como resposta a estes problemas, e recebendo o impulso das aproximações estatísticas utilizadas para resolver tarefas de EI e da crescente disponibilidade de poder de processamento, deu-se uma alteração de paradigma. Sistemas de PLN que seguiam metodologias orientadas para a informação ou baseadas em Aprendizagem Automática (AA) utilizando corpus, agora bastante mais acessíveis, tornaram-se novamente o grande foco nos anos 90 [8, 27]. Atualmente existe um grande movimento em torno de técnicas de aprendizagem não supervisionada ou parcialmente supervisionada. Esta aproximação permite não só tirar partido de corpus revistos como também do vasto conteúdo espalhado pela Internet para alimentar modelos estatísticos e aplicações baseadas em AA [42, 2, 15].

2.1.3 Corpus Linguístico

Um corpus pode ser definido como uma coleção de textos escritos ou falados que se assume serem representativos de uma linguagem e que foram recolhidos para serem utilizados para

análise. Normalmente assume-se que o conteúdo do corpus ocorreu naturalmente, que foi recolhido de acordo com critérios e propósitos bem definidos e que este é representativo de blocos maiores da linguagem, selecionados de acordo com uma tipologia [45].

Um corpus pode ser utilizado com vários fins, como por exemplo:

- Análise estatística e teste de hipóteses;
- Validação de regras da língua;
- Verificação de ocorrências de palavras, expressões ou pontuação.

Alguns corpus incluem informação adicional com o objetivo de facilitar o seu estudo e acrescentando valor ao mesmo. Para incluir esta informação adicional estes podem ser submetidos a um processo a que se dá o nome de anotação. São comuns anotações relativas à morfologia, à sintaxe e à semântica. A existência destas anotações gera no entanto algumas limitações no que diz respeito à criação e manutenção do corpus. Assegurar a consistência e completude de um corpus anotado requer bastante trabalho, o que leva a que estes tenham em geral menor dimensão, tornando menos significativas as conclusões retiradas para o domínio que representam.

2.1.4 Léxico e Léxico de Sentimentos

Na Linguística, um léxico é o conjunto aberto de todas as palavras e elementos morfológicos com significado possíveis numa Língua [1]. Pode ser descrito como uma das componentes de uma linguagem, sendo a outra a componente a gramática onde se encontram descritas todas as regras para uma correta construção frásica. O léxico contém todos os símbolos que podem ser utilizados na linguagem, vulgarmente referido como vocabulário neste contexto. Dependendo um pouco do âmbito, é mutável, encontrando-se normalmente sujeito a um processo evolutivo que acompanha a língua.

Um léxico de sentimentos é, no contexto da Linguística e do PLN, um léxico no qual foi incluída informação sobre a polaridade (positiva, negativa, neutra e possivelmente outras) estimada para cada elemento constituinte. Pode incluir também indicações adicionais para casos em que a polaridade de um elemento varia em função do contexto em que surge. São exemplos de léxicos de sentimentos o Sentilex-PT [43] e o SentiWordNet [17] (figuras 2.3 e 2.4).

2.1.5 Análise Morfossintática

O termo sintaxe deriva do Grego Antigo "*sýntaxis*", um nome verbal que literalmente significa "combinação" ou "disposição". Tradicionalmente refere-se ao ramo da gramática que lida com o modo como as palavras, corretamente flexionadas ou não, são dispostas

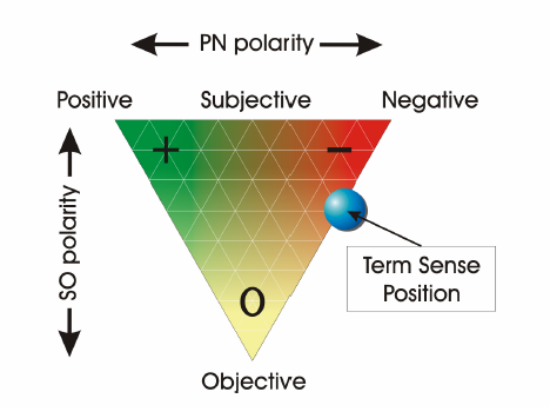

Figura 2.3: Representação gráfica adotada pelo SentiWordNet para representar a opinião do sentido de um termo. [17]

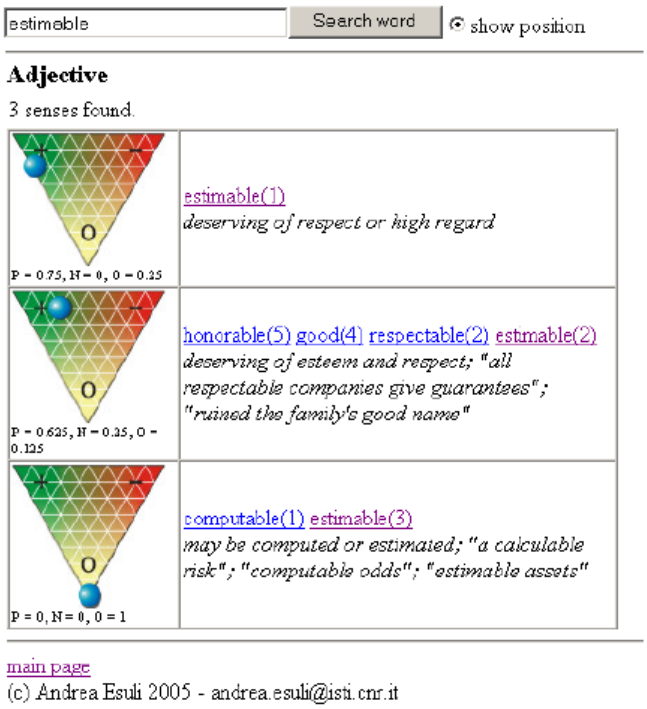

Figura 2.4: Visualização da opinião sobre o termo "estimável" (*estimable*) no SentiWordNet. [17]

ao ao longo da frase para obter um significado [29]. A sintaxe define assim o modo como devem ser construídas e estruturadas as frases de uma determinada linguagem. Esta pode variar consoante a língua em fatores por vezes óbvios mas significativos, como a ordem dos elementos principais - verbo, sujeito e complementos [49].

O termo morfologia é geralmente atribuído ao poeta, romancista, dramaturgo e filósofo

Alemão Johann Wolfgang von Goethe (1749-1832) que o utilizou no início do século XIX no âmbito da biologia. A palavra tem origem do Grego *morph-* cujo significado é "forma", a morfologia é precisamente o estudo da forma ou formas. Na Linguística a morfologia refere-se ao processo de formação de palavras e à sua estrutura, flexão e classe gramatical [4].

O conceito de morfossintaxe surge assim englobando os anteriores. A analise morfossintática consiste em apreciar conjuntamente a sintaxe - disposição na frase - e morfologia - flexão e classe gramatical - das palavras de uma frase.

2.1.6 Reconhecimento de Entidades Mencionadas

O termo "Entidade Mencionada", hoje em dia bastante utilizado no âmbito do PLN, foi inicialmente introduzido na 6ª *Message Understanding Conference*[6] (MUC-6) que ocorreu em Novembro de 1996. O grande foco da MUC foi na altura a EI, onde informação estruturada relativa a atividades de empresas e atividades de defesa é extraída de fontes textuais. Ao definir esta tarefa chegou-se à conclusão que é fundamental reconhecer unidades de informação como os nomes, incluindo nomes de pessoas, organizações e locais e expressões numéricas como a hora, data, dinheiro e percentagens. Reconheceu-se que identificar referencias a estas entidades é uma das tarefas importantes da EI, e foi então baptizada de "Reconhecimento de Entidades Mencionadas" (REM) [34].

O idioma é um dos fatores de influência no REM, sendo que grande parte da investigação é feita na Língua Inglesa, existe paralelamente um grande interesse para com a independência de linguagem e multilíngua. O género dos textos (notícias, científico informal, etc.) e o domínio (jardinagem, desporto, negócios, etc.) são também fatores com impacto nos sistemas de REM. No entanto, estes têm recebido menos atenção tendo em conta que são muito poucos os estudos realizados que abordam estas condicionantes. O trabalho inicial formulou o problema de REM como o reconhecimento de nomes próprios de um modo geral. Os tipos mais estudados são normalmente as três especializações dos nomes próprios: nomes de pessoas, locais e organizações. Estes são colectivamente conhecidos como "*enamex*" desde a MUC-6 e são por vezes divididos em várias subcategorias. O tipo "*miscellaneous*" é usado para referenciar entidades que saem fora dos *enamex*. "*Timex*" é o termo por vezes utilizado para aglomerar os grupos de tempo, data e "*numex*" para dinheiro e percentagens. Trabalhos mais recentes procuram não limitar os tipos possíveis e são referenciados como sendo de domínio livre e tentam utilizar a maioria dos tipos de nomes mais frequentes. A habilidade de reconhecer entidades previamente desconhecidas é uma parte essencial dos sistemas de REM. Esta capacidade assenta sobre o reconhecimento e regras de classificação acionadas pelas características distintas associadas a exemplos corretos e incorretos. Enquanto estudos anteriores eram maioritariamente baseados em regras definidas manualmente, os mais recentes utilizam aprendizagem supervisionada como método de induzir automaticamente sistemas baseados em regras ou algoritmos se-

[6] http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html

quenciais tomando como base alguns exemplos de treino. Tipicamente a aprendizagem supervisionada consiste num sistema que consome um corpus anotado, memoriza listas de entidades e gera regras de desambiguação baseadas nas características mais dominantes. Quando exemplos de treino não estão disponíveis, os sistemas baseados em regras manuais continuam a ser a técnica preferencial. Os sistemas parcialmente supervisionados são relativamente recentes e assentam numa técnica a que se dá o nome de *bootstrapping*. Esta técnica envolve um pequeno grau de supervisão, por exemplo um conjunto de *seeds* ou "sementes" para iniciar o processo. Neste contexto, uma semente poderá ser uma regra base fornecida ao sistema. Este utiliza esse conjunto de regras base, as sementes, para encontrar os primeiros resultados. Destes resultados são inferidas novas regras e adicionadas ao conjunto já existente. O processo é repetido iterativamente com um número de sementes cada vez maior permitindo assim aumentar progressivamente os resultados obtidos. Nos métodos não supervisionados o processo é completamente automatizado. Tipicamente as técnicas são baseadas na utilização de recursos lexicais, como por exemplo o WordNet [31], padrões lexicais e estatísticas obtidas a partir de grandes corpus não anotados [34].

2.2 Ferramentas

Durante o período de investigação, e posteriormente durante a construção do protótipo que fundamenta a escrita desta dissertação, foram estudadas e utilizadas algumas ferramentas importantes. Neste ponto descrevem-se essas ferramentas procurando facilitar a compreensão das referências às mesmas nos capítulos que se seguem.

2.2.1 LX-Tagger

O LX-Tagger é uma ferramenta desenvolvida pelo NLX-Group[7] na Universidade de Lisboa. Esta ferramenta desempenha as funções de analisador morfossintático, já referidas no ponto 2.1.5, atribuindo a cada palavra a etiqueta correspondente à sua classe. No seu desenvolvimento foi utilizada uma outra ferramenta, o MXPOST[37], e um corpus anotado manualmente composto por cerca de 600.000 termos [7].

```
1  O/DA Socrates/PNM tem/VAUX estado/PPT horrível/ADJ .*//PNT
```

Listagem 1: Resultado de processamento do LX-Tagger.

A listagem 1 mostra um exemplo de uma frase processada pelo LX-Tagger. A ferramenta é disponibilizada para descarga ou utilização via serviço[8]. No entanto a versão descarregada não possui toda a funcionalidade da sua versão serviço.

[7] Grupo de Fala e Linguagem Natural do Departamento de Informática da Universidade de Lisboa, Faculdade de Ciências.

[8] Mais informação em http://lxcenter.di.fc.ul.pt/tools/pt/LXTaggerPT.html.

2.2.2 PALAVRAS

O PALAVRAS é um analisador morfossintático para Português desenvolvido por Eckhard Bick no âmbito de um projeto de doutoramento (1994-200) na Universidade de Aarhus na Dinamarca. Este baseia-se num léxico de cerca de 50.000 lemas e milhares de regras gramaticais para fornecer uma análise sintáctica e morfológica de qualquer texto. Seguindo uma metodologia de *Constraint Grammar*[9] e utilizando um conjunto de etiquetas bastante diversificadas, a ferramenta atinge níveis de correção de 99% na classificação morfológica e de 97-98% na sintaxe [6].

```
1  EXC:fcl
2  =SUBJ:np
3  ==>N:art('o' <artd> M S)  O
4  ==H:prop('Socrates' M/F S)  Socrates
5  =P:vp
6  ==AUX:v-fin('ter' PR 3S IND)  tem
7  ==MV:v-pcp('estar' M S) estado
8  =SC:adj('horrível' M/F S) horrível
9  =.
```

Listagem 2: Resultado de processamento do PALAVRAS.

A listagem 2 mostra um exemplo do processamento realizado pelo PALAVRAS. Atualmente é possível fazer uso deste analisador através do projeto VISL[10].

2.2.3 Weka

O Weka consiste numa coleção de algoritmos implementados na linguagem de programação JAVA, com o propósito de serem utilizados em tarefas de mineração de dados. A sua utilização pode ser feita através da interface própria ou embebendo as suas funcionalidades na nossa aplicação. A ferramenta possui a capacidade de realizar pré-processamento, classificação, regressão, *clustering*, associação e visualização de dados. É especialmente adequado para tarefas de AA, incluindo uma quantidade considerável de algoritmos implementados para o efeito [22].

O Weka é software livre e encontra-se disponível para utilização sobre a licença GNU General Public License[11].

[9] Paradigma metódico de PLN introduzido por Fred Karlsson (Universidade de Helsínquia, Finlândia) em 1992.

[10] Visual Interactive Language Learning, http://visl.sdu.dk.

[11] http://www.gnu.org/licenses/gpl.html

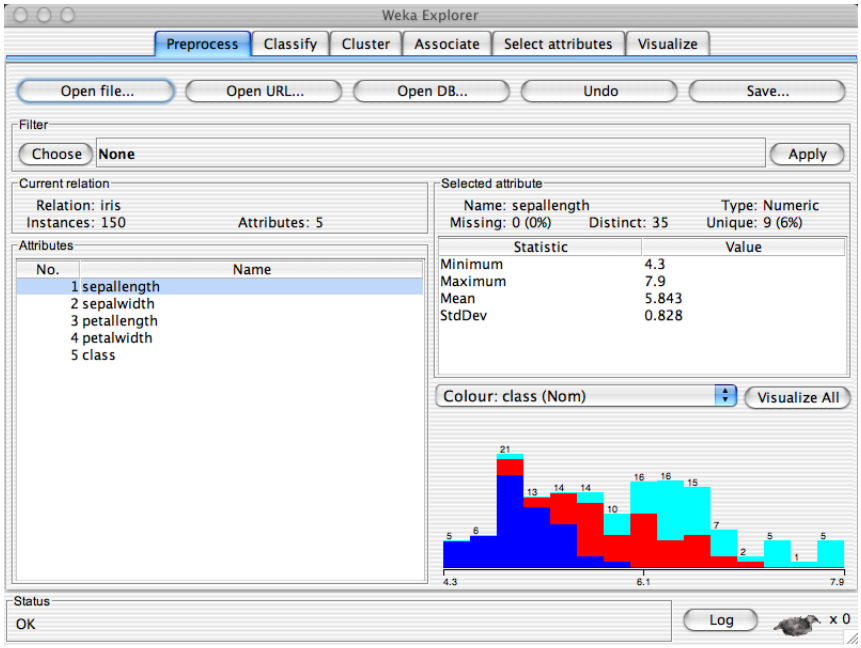

Figura 2.5: Interface gráfica do Weka.

2.3 Conclusões

Ao longo deste capítulo foram apresentados alguns conceitos valiosos para a compreensão do assunto que trata esta dissertação. Adicionalmente, foram também descritas ferramentas externas, que contribuíram de alguma forma na investigação e no desenvolvimento deste projeto. O seu propósito ficará mais claro ao longo dos restantes capítulos, à medida que vão sendo mencionadas no contexto da sua utilização.

Capítulo 3

Estado da Arte

Com este capítulo pretende-se ilustrar o estado atual dos sistemas de AS. É realizado um pequeno resumo das origens do tema, apontando algumas das razões pelas quais tem captado a atenção da comunidade científica. Finalmente são apresentados e analisados alguns projetos que exploram as metodologias tidas como mais comuns.

3.1 Trabalhos Iniciais

Na ultima década a importância atribuída à AS cresceu significativamente. Porém, não se pode dizer que a ideia seja algo assim tão recente. No final dos anos 70 Carbonell elaborou já um trabalho abordando a temática da deteção de subjetividade [9]. Nas décadas de 80 e 90 a investigação continuou, principalmente incidindo sobre temas como a interpretação de metáforas, compreensão da narrativa, deteção de ponto de vista e mais uma vez, deteção de subjetividade. O ano de 2001 parece ser o ano em que se consolidou finalmente a importância da AS [35].

O crescimento da *World Wide Web* e a facilidade de partilha de conteúdos *online*, rapidamente torna evidente a necessidade de explorar este novo recurso. A evolução da capacidade de processamento do *hardware* e crescente disponibilidade de corpus leva ao surgimento de novos algoritmos de AA, outrora impossíveis de implementar ou sem a capacidade de produzir os resultados necessários. O enorme desafio intelectual que a área representa atrai investigadores, sendo que nos últimos anos centenas de artigos foram publicados diretamente relacionados com o tema. Ao mesmo tempo, as pressões da indústria, principalmente de áreas ligada à IE, tornam-se cada vez mais fortes exprimindo a necessidade de ferramentas que permitam saber "*o que outros pensam*" [35].

3.2 Aprendizagem Automática

A AA é caracterizada por um conjunto de técnicas cuja base passa geralmente por uma fase obrigatória de treino, sendo necessário facultar dados de natureza empírica ao classificador. O objetivo passa então por utilizar o conhecimento adquirido no treino para tomar decisões inteligentes em situações não encontradas previamente. Os algoritmos podem ser categorizados relativamente ao nível de interação manual necessário na fase de aprendizagem.

A aprendizagem supervisionada consiste, neste contexto, em fornecer ao classificador toda informação necessária ao seu treino devidamente catalogada e revista manualmente. No caso de uma aprendizagem não supervisionada, a informação de treino não sofre qualquer revisão manual, o classificador terá de inferir quais as características relevantes e agrupar a informação de acordo com estas, tentando obter resultados significativos. Finalmente, alguns algoritmos surgem como uma aproximação intermédia, os parcialmente supervisionados, utilizando como *input* não só informação revista como também não revista. Normalmente a informação não revista é em maior quantidade. A aproximação pode passar por inferir a revisão da informação não revista com base na já revista, aumentado assim a base de treino de forma automática [41].

3.2.1 Sentiment140

O atual Sentiment140[1], previamente denominado de Twitter Sentiment, consiste num sistema de classificação de mensagens oriundas do serviço de *microblogging* Twitter. O objetivo deste sistema é classificar *tweets* como positivas ou negativas em relação a um termo, sobre o qual se realiza uma pesquisa. Como resultado são disponibilizados gráficos indicando o sentimento predominante e exemplos de mensagens coloridas de acordo com a classificação atribuída. A figura 3.1 ilustra uma pesquisa realizada no sistema. Atualmente é possível utilizar o sistema em duas línguas, Inglês e Espanhol.

O Sentiment140 foi desenvolvido por Alec Go, Richa Bhayani e Lei Huang, três estudantes de pós-graduação na Universidade de Stanford. O seu trabalho surge numa altura em que a investigação relacionada com a classificação de mensagens provenientes de *microblogs* praticamente ainda não existia. A motivação para desenvolvimento do sistema residiu principalmente na variedade de domínio, em oposição a outras fontes onde se trata apenas um conteúdo especifico, na rede Twitter qualquer assunto é valido. A facilidade na obtenção de mensagens da rede foi também um fator apelativo, através da API disponibilizada pelo Twitter é possível recolher milhões de *tweets* sem dificuldade [20].

Na construção do sistema foi utilizado um corpus de cerca de 1.800.000 *tweets* automa-

[1]Disponível em http://www.sentiment140.com.

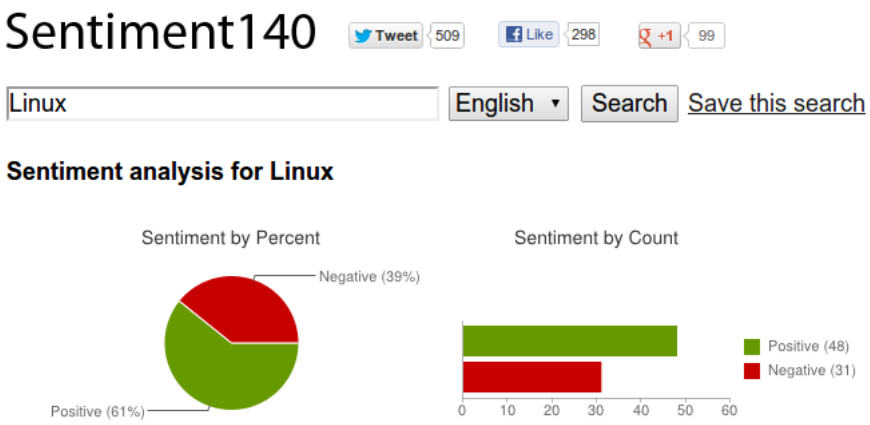

Figura 3.1: Resultado de pesquisa realizada no Sentiment140 pelo termo "Linux".

ticamente classificadas utilizando Supervisão Distante[2], recorrendo a *emoticons*[3] para o efeito. Apenas foram consideradas duas classes na classificação, positivo e negativo. Esta limitação é justificada pelo facto de que a maioria das mensagens são efetivamente neutras, não adicionando significância aos resultados [20].

Características	**Keyword**	**Naive Bayes**	**MaxEnt**	**SVM**
Unigram	65.2	81.3	80.5	82.2
Bigram	N/A	81.6	79.1	78.8
Unigram+Bigram	N/A	82.7	83.0	81.6
Unigram+POS	N/A	79.9	79.9	81.9

Tabela 3.1: Avaliação do Sentiment140, medidas para a Precisão [20].

Os métodos de classificação escolhidos na avaliação foram *Keyword*[4], Naive Bayes[5], Maximum Entropy[6] (MaxEnt), e Support Vector Machines[7] (SVM). As características utilizadas na aprendizagem foram *unigrams*, *bigrams*, *unigrams* e *bigrams* e *unigrams*[8] com informação morfológica. Na fase de treino foram removidos os *emoticons* devido ao impacto negativo que representaram. Com exceção do método de Naive Bayes, os outros algoritmos foram negativamente influenciados pela presença destes elementos. Esta situação sucede devido aos modelos matemáticos que suportam o MaxEnt e o SVM. Foi ainda realizado um pré-processamento onde foram removidos *links*, nomes de utilizador e repetições de letras, elementos comuns à maioria das mensagens do género. Os resultados

[2]Um algoritmo de Supervisão Distante é um algoritmo semi-supervisionado que recorre a uma ou mais características dos dados, assumidas como identificadoras da classe, para realizar a classificação de forma automática.

[3]Representação de uma expressão facial utilizando sinais de pontuação, números e letras.

[4]Conjuntos de palavras chave representativas de cada uma das classes em classificação.

[5]Método de classificação baseado no teorema de Bayes.

[6]Método estatístico de classificação baseado no princípio da maximização da entropia.

[7]Método de aprendizagem supervisionada que dado um input atribui uma classificação binária.

[8]Um *n-gram* é uma sequência contínua de n elementos numa frase ou texto. São casos particulares os *unigrams*, *bigrams* e *trigrams* que correspondem a *n-grams* de tamanho 1, 2 e 3 respectivamente.

obtidos durante o desenvolvimento do sistema são listados na tabela 3.1.

No momento da escrita desta dissertação a documentação disponibilizada no *site* oficial[9] do sistema indica que o classificador utilizado é o MaxEnt. No mesmo local são também indicados alguns dos problemas ainda não abordados pela ferramenta, dos quais se destacam:

Deteção de Subjetividade Não é realizada uma deteção de subjetividade por forma a evitar a classificação de *tweets* isentas de sentimento.

Negação A deteção de negação é algo complexa tendo em conta a abordagem escolhida para o sistema.

Comparação Não são detetados casos em que ocorre comparação entre termos, resultando numa avaliação semelhante para ambos.

Sarcasmo Identificação de sarcasmo é uma das problemáticas atuais da AS.

Classificação Morfológica A utilização de um analisador morfossintático treinado em corpus de um domínio completamente diferente gera erros frequentes de classificação.

3.2.2 *Target-dependent Twitter Sentiment Classification*

Target-dependent Twitter Sentiment Classification tem origem no trabalho realizado por Jiang et al. [26]. Este sistema foi criado com a ambição de colmatar algumas das falhas existentes nas aproximações atuais, de entre as quais consta o Sentiment140 já aqui apresentado.

Tal como o Sentiment140, o objetivo é analisar texto proveniente do serviço de *microblogging* Twitter, classificando-o de acordo com o sentimento direcionado ao termo pesquisado. Ao analisar os resultados do Sentiment140, procurando identificar as suas falhas, identificou-se que grande parte dos erros de classificação ocorriam em duas situações [26].

1. *People everywhere love Windows & vista. Bill Gates*
2. *Windows 7 is much better than Vista!*

Na primeira situação, o sentimento seria atribuído ao alvo "*Bill Gates*", sendo que na realidade não existe qualquer opinião sobre esse alvo no texto. No segundo caso, a classificação é positiva para qualquer um dos alvos ("*Windows 7*" e "*Vista*"), o que também não está de todo correto. Através dos testes realizados, Jiang et al. identificaram que cerca de 40% dos erros de classificação cometidos pelo Sentiment140 eram devido a problemas semelhantes aos dois aqui apresentados. Procurando solucionar esta situação o seu sistema procura explorar dois conceitos:

[9] http://help.sentiment140.com

Dependência de Alvo Estabelecer uma ligação entre o sentimento na frase e o seu alvo.

Noção de Contexto Utilizar as relações entre *tweets* para estabelecer um contexto.

À semelhança da maioria dos sistemas atuais com o mesmo propósito, a aproximação escolhida é baseada em AA. Porém, tipicamente nestes sistemas as características de classificação são completamente independentes do alvo, levando assim a um dos problemas já mencionados. Procurando contornar esta situação os autores recorreram àquilo a que deram o nome de *target-dependent features*[10]. Estas características são geradas recorrendo a informação obtida através da analise sintática do texto, permitindo fazer a aprendizagem levando em conta o posicionamento do alvo em relação ao sentimento [26].

Por definição, as mensagens num serviço de *microblogging* são bastante pequenas. Este fator complica a perceção do tipo de sentimento que incide sobre o alvo. De forma a melhorar a performance nestes casos, foi introduzida uma noção de contexto procurando estabelecer uma relação entre mensagens. Para definir este contexto, recorrendo a informação disponibilizada pelo Twitter, são consideradas *tweets* colocadas pelo mesmo utilizador, na mesma conversação, em resposta ou que respondem à *tweet* em análise.

Em termos de metodologia, a classificação de uma *tweet* pelo sistema passa por três fases distintas:

1. **Deteção de Subjetividade** Por forma a eliminar a priori *tweets* objetivas é realizada uma análise de subjetividade, atribuindo a classe neutra a mensagens identificadas como objetivas. Aquelas que são identificadas como subjetivas passam para a fase seguinte do processo.

2. **Deteção de Polaridade** Nesta fase é atribuída uma das restantes classes à mensagem, classificando o sentimento para com o alvo como positivo ou negativo.

3. **Deteção de Contexto** É realizada uma deteção do contexto da mensagem, recorrendo a uma aproximação baseada em grafos. Aqui são exploradas as relações entre *tweets* e utilizado o sentimento atribuído ao alvo nessas relações, considerando essa informação na avaliação final.

Para os pontos 1 e 2 foram construídos classificadores SVM. Como características do classificador de polaridade foram utilizadas as palavras, pontuação, *emoticons*, *hashtags*[11] e informação sobre polaridade. Esta ultima consiste numa contagem de quantas palavras positivas ou negativas constam na mensagem, utilizando um léxico pré definido para o efeito. Em adição a estas são consideradas também as características definidas como *target-dependent*, já mencionadas.

[10] Características dependentes do alvo.
[11] Palavra ou frase que contém como prefixo o símbolo "#".

	Medida-F1 (%)		
Precisão (%)	pos	neu	neg
68.3	63.5	71.0	68.5

Tabela 3.2: Avaliação do *Target-dependent Twitter Sentiment Classification*, Precisão e Medida-F1 por classe [26].

Para o ponto 3, para além das relações entre *tweets*, foi considerada uma aproximação a que os autores deram o nome de *extended targets*[12]. Esta explora a existência de relações entre entidades, permitindo que o sentimento expresso sobre uma entidade relacionada com o alvo principal influencie a classificação deste. Foram consideradas extensões coisas como os produtos de uma companhia, sendo a companhia o alvo principal, ou as características de determinado produto tendo o produto como alvo.

Existe também um pré-processamento realizado às mensagens onde é feita uma normalização, uma análise morfossintática e *stemming*[13]. A tabela 3.2 contém os resultados finais obtidos pelo sistema. Estes resultados referem-se à identificação correta do sentimento e respetiva associação ao alvo, razão pela qual os resultados podem parecer inferiores aos do Sentiment140.

3.3 Orientação Semântica

A Orientação Semântica (OS) é uma metodologia bastante popular na solução de problemas de EI e PLN. Por vezes denominados de sistemas de regras ou baseados em regras, o seu funcionamento apoia-se no reconhecimento de padrões ou na utilização de algum tipo de indicadores previamente definidos. Este tipo de aproximação é considerada como não supervisionada, visto não ser necessária a categorização manual de conteúdo para treino, não se aplicando inclusivamente o conceito de treino [35].

Em alguns casos, os sistemas que seguem esta metodologia fazem uso de recursos auxiliares, sendo comum a utilização de léxicos de sentimento ou dicionários de sinónimos. O processo de classificação consiste geralmente na aplicação de um algoritmo, que utiliza os recursos já mencionados, e recorre a um conjunto de regras pré-definidas para atribuição da classe.

3.3.1 SentiCorr: *Multilingual Sentiment Analysis of Personal Correspondence*

O SentiCorr é um sistema de análise automática de sentimento desenhado para classificar texto proveniente de redes sociais e *e-mail*. O objetivo do sistema, segundo os autores, é dar uma perceção aos utilizadores da quantidade de texto com carácter positivo e negativo que estes escrevem e lêem no seu dia-a-dia.

[12] Extensão de alvo.
[13] Processo de redução de uma palavra à sua forma base ou *stem*.

Para facilitar a utilização do sistema na classificação de *e-mails* foi criado um *plugin* para o Microsoft Outlook, que destaca quais as mensagens positivas e negativas. As classificações efetuadas pelo SentiCorr são armazenadas numa Base de Dados, permitindo uma posterior análise ao longo de dimensões como o tempo, emissores, recetores e outras semelhantes [48].

O sistema foi desenhado de forma modular, sendo que uma das suas propriedades é ser multilingue, suportando atualmente Inglês e Holandês. Na classificação é atribuída uma de três classes possíveis: positiva, negativa e objetiva. O processo compreende quatro fases ilustradas na figura 3.2 e descritas de seguida.

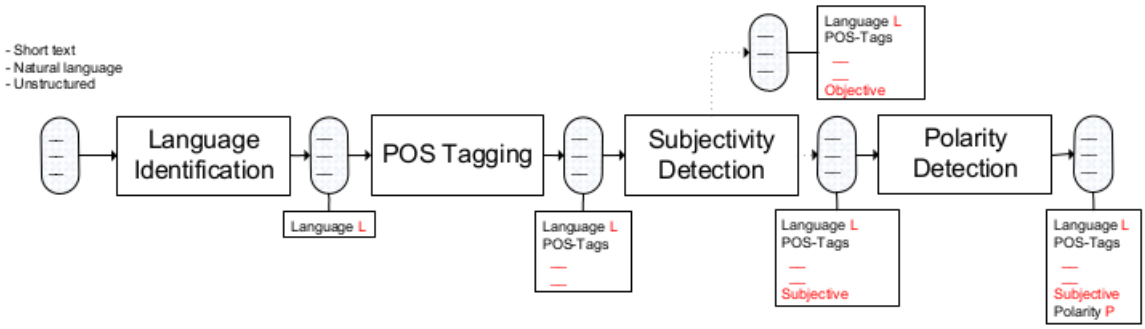

Figura 3.2: Processo de classificação do SentiCorr [48].

1. **Identificação da Língua** A deteção de língua é levada a cabo recorrendo ao algoritmo LIGA [47].

2. **Análise Morfossintática** Neste passo é utilizado o analisador TreeTagger [40] que possui modelos para várias línguas disponíveis publicamente.

3. **Deteção de Subjetividade** Para proceder a uma seleção das mensagens subjetivas recorreu-se à ferramenta AdaBoost [18].

4. **Deteção de Polaridade** Finalmente é detetada a polaridade dos elementos identificados como subjetivos no ponto anterior. Neste passo foi introduzido um algoritmo de deteção de polaridade a que se deu o nome de Rule Based Emission Model (RBEM).

O RBEM baseia-se na premissa de que qualquer elemento de uma mensagem pode emitir sentimento positivo ou negativo. Para identificar esse sentimento foram criadas regras definidas através de oito padrões: positivo, negativo, amplificador, atenuador, inversor, continuador e *stop*. A deteção de polaridade resulta da identificação de que padrões se encaixam no texto a classificar, calculando posteriormente o valor final que dita qual a classe a atribuir [46, 48].

O sistema realiza uma classificação ao nível da frase. Esta granularidade é justificada pela suposição de que uma frase retirada de uma rede social possui apenas um sentimento, principalmente devido ao seu tamanho reduzido. E que no caso de existir mais que um sentimento na mesma frase, existe sempre um que se pode considerar o predominante.

O sistema foi avaliando com textos retirados do Twitter, Facebook e Hyves[14] ficando a performance deste bastante perto dos 70%, na classificação das três classes [48].

3.3.2 PIRPO: *An Algorithm to deal with Polarity in Portuguese Online Reviews from the Accommodation Sector*

Existem vários *sites* que permitem aos seus utilizadores deixar opiniões sobre hotéis, restaurantes e serviços semelhantes dos quais usufruíram. A utilidade desta informação para as empresas ligadas ao ramo é reconhecida, no entanto os custos associados ao seu tratamento manual são deveras elevados. É neste alinhamento que surge o projeto PIRPO, com o objetivo de criar uma ferramenta capaz de analisar automaticamente o sentimento contido nos comentários deixados pelos utilizadores.

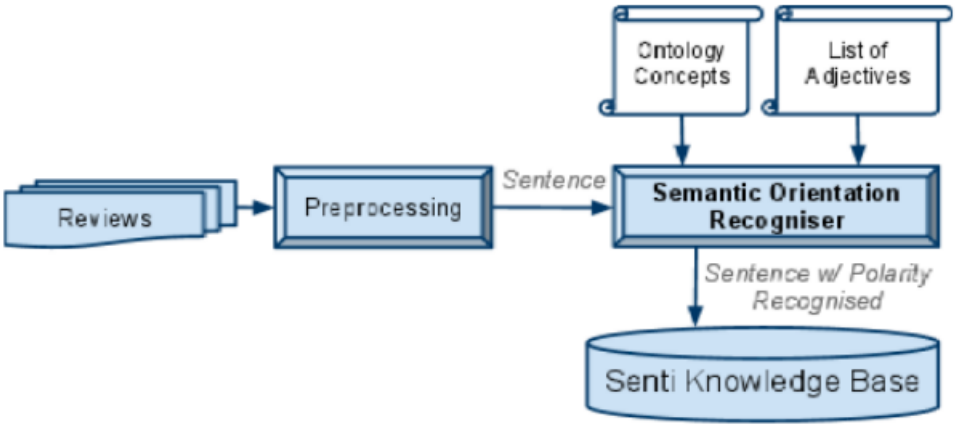

Figura 3.3: Arquitetura geral do PIRPO [11].

Em termos de funcionamento, os comentários a analisar são alvo de um pré-processamento inicial. Nesta fase são divididos em frases e detetados aqueles que foram inseridos como positivos e negativos. Esta última operação resulta do facto de alguns *sites* de partilha de opinião forçarem esta separação durante o processo de submissão. O sistema faz uso de dois recursos auxiliares para realizar a classificação, descritos de seguida [11].

Hontology Ontologia multilingue de conceitos para o domínio da hotelaria descrita por Chaves et al. [12].

[14] Rede social holandesa.

Léxico de Sentimento Lista de adjetivos categorizados como positivos (1), negativos (-1) ou neutros (0) descrita por Souza et al. [44].

O módulo principal, designado por *Semantic Orientation Recognizer*, procura reconhecer em cada frase, sinais de sentimento para com os conceitos presentes no *Hontology*. Para cada um dos conceitos encontrados, é realizada uma busca numa janela de palavras definida em torno do conceito, procurando encontrar palavras presentes no léxico de sentimento. A soma dos vários indicadores de polaridade encontrados é então normalizada, representando o sentimento que incide sobre o conceito. O resultado da classificação efetuada é armazenado numa base de conhecimento, permitindo posterior consulta e análise. A arquitetura do sistema é ilustrada na figura 3.3.

	Prec			**Cob**			**F1**		
Conceitos	P	N	M	P	N	M	P	N	M
Hotel	0.20	0.00	1.00	0.33	0.00	0.22	0.25	0.00	0.36
Localização	0.00	0.00	0.89	0.00	0.00	0.24	0.00	0.00	0.37
Quarto	0.11	0.00	1.00	1.00	0.00	0.19	0.20	0.00	0.32
Staff	0.08	0.00	1.00	1.00	0.00	0.12	0.15	0.00	0.21
Média	0.10	0.00	0.97	0.58	0.00	0.19	0.15	0.00	0.32

Tabela 3.3: Resultados da avaliação preliminar do PIRPO. Valores para a Precisão, Cobertura e Medida-F1 por classe e por conceito [11].

Na avaliação do sistema foram utilizadas opiniões retiradas do Tripadvisor[15] e do Booking[16], perfazendo um total de 180 comentários em Português. Este conjunto foi sujeito a um processo de classificação manual que categorizou as várias opiniões em três classes: positiva, negativa e mista [11]. Estas são as mesmas classes atribuídas pelo sistema como resultado do seu processamento. A tabela 3.3 apresenta os resultados preliminares obtidos.

3.4 Conclusões

A AS é uma temática algo recente, embora os primeiros passos tenham sido dados no inicio da década de 80, só nos últimos anos esta tem merecido uma maior atenção. O crescente interesse pela matéria surge como resultado da pressão de dois mundos: da comunidade científica, pelo enorme desafio intelectual que proporciona, e do mundo empresarial, pelo valor de saber o que os clientes realmente querem. Em paralelo, a tecnologia possui finalmente a capacidade de concretizar ideias que há alguns anos eram tidas como impossíveis, promovendo o aparecimento de novos métodos.

Atualmente existem duas abordagens, consideradas como mais comuns, no desenho de sistemas de AS: os sistemas baseados em métodos de AA e os sistemas que recorrem a regras ou de orientação semântica. Foram apresentados alguns projetos que exploram

[15] http://www.tripadvisor.com
[16] http://www.booking.com/

ambos os paradigmas, e que ilustram o estado atual da área.

Capítulo 4

Solução Proposta

Neste capítulo é descrita a abordagem seguida no desenvolvimento do protótipo que permitiu a recolha de resultados. É realizada uma descrição geral do funcionamento de todo o sistema e, mais aprofundadamente, de cada um dos módulos que o constitui.

4.1 Arquitetura Geral

O protótipo desenvolvido é composto por três módulos principais. Cada um destes módulos procura responder a uma problemática do PLN de forma a atingir os resultados a que esta dissertação se propõe.

Analisador Morfossintático Este módulo é responsável por classificar morfológica e sintaticamente o texto. O seu processamento é fundamental para o funcionamento das restantes componentes.

Reconhecedor de Entidades Mencionadas Neste módulo são identificadas as entidades sobre as quais incide o sentimento. Sem esta informação não é possível identificar o objeto da opinião.

Analisador de Sentimentos O módulo final é responsável por atribuir a classificação de sentimento, utilizando para o efeito informação recebida dos módulos anteriores.

A figura 4.1 mostra uma perspectiva superficial da interação entre os vários módulos que compõem o sistema. De forma a manter uma estrutura modular e a independência tecnológica entre as várias componentes, foi necessário escolher uma forma de comunicação.

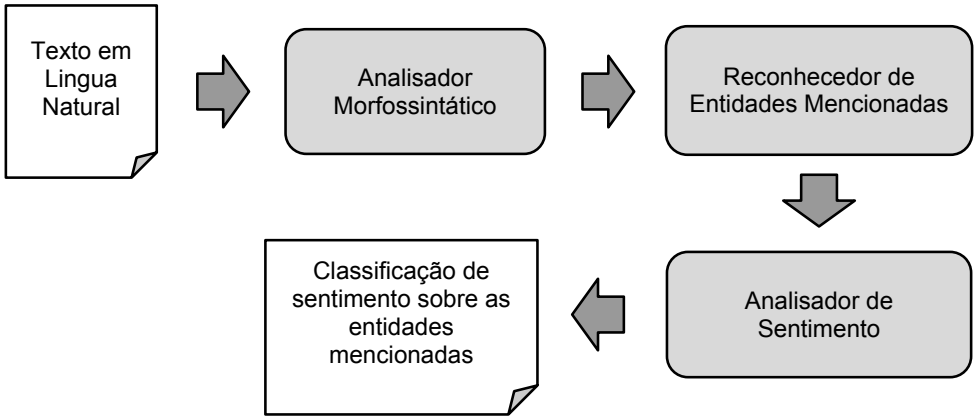

Figura 4.1: Arquitetura geral do sistema.

A opção escolhida foi a troca de mensagens no formato XML[1]. Nos próximos pontos são descritos cada um dos módulos em detalhe e fornecidos exemplos das mensagens trocadas.

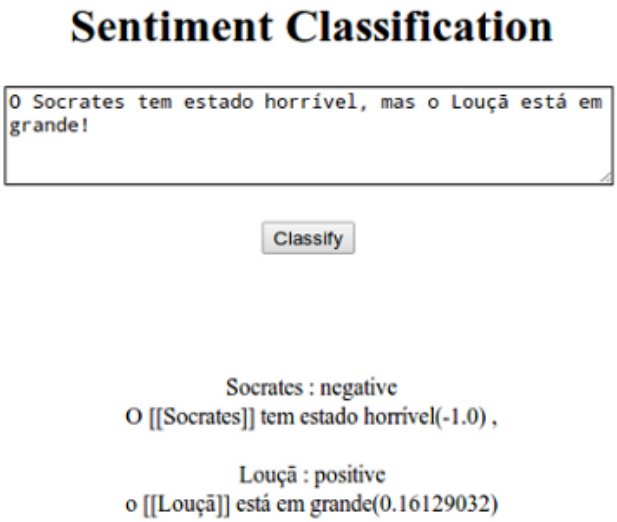

Figura 4.2: Interface simples do sistema.

Por forma a visualizar resultados de forma simples, foi criada uma pequena interface *web*. Esta permite questionar o sistema de um modo mais amigável relativamente ao sentimento de um texto. A figura 4.2 ilustra esta interface.

[1] Acrónimo para *eXtensible Markup Language*.

4.2 Analisador Morfossintático

Este componente foi tido como necessário para o desenvolvimento do sistema por se compreender serem relevantes as propriedades resultantes do seu processamento. Os conceitos inerentes ao seu funcionamento já foram enunciados no ponto 2.1.5. Descreve-se aqui o seu propósito no protótipo desenvolvido e a forma como opera com as restantes componentes.

4.2.1 Finalidade

A informação recolhida por este módulo é utilizada pelos restantes componentes de várias formas. Enumeram-se as principais de que o sistema beneficiou.

- Os substantivos são a classe gramatical associada aos nomes. Estes podem ser comuns ou próprios, fornecendo uma boa indicação de onde se encontra o sujeito de uma frase. Estes factos são tidos em conta pelo módulo de REM para detetar o objeto de opinião.

- A existência de adjetivos no texto é uma forte indicação da existência de subjetividade. É feito uso desta característica em conjunto com a conjugação dos verbos para categorizar o tipo de opinião presente.

- A sintaxe de uma frase indica-nos como estão dispostas as suas várias componentes, permitindo assim estabelecer algumas regras relativamente à abrangência. A oração foi uma das componentes sintáticas exploradas para isolar a influência de elementos com polaridade, evitando assim a propagação incorreta de um sentimento.

- Os advérbios, têm a capacidade de alterar ou inverter o sentimento através da sua influência sobre o verbo. Esta propriedade é tida em conta no módulo de AS.

De acordo com fluxo do sistema ilustrado na figura 4.1, o resultado do processamento deste módulo alimenta o módulo de REM. Como tal, o formato de saída corresponde ao formato de entrada do módulo seguinte.

4.2.2 Evolução

A componente de AM passou por duas versões distintas durante o desenvolvimento do sistema. Para além de alterações no formato das mensagens a passar ao módulo seguinte, a utilização de ferramentas foi também uma variante. Neste ponto descreve-se a sua evolução.

AM v1.0

A primeira versão desta componente utilizou como ferramenta base o LX-Tagger, já descrito no ponto 2.2.1. Esta ferramenta, na forma em que é disponibilizada, apenas realiza uma análise morfológica do texto. Esta versão do componente restringe-se assim apenas a este tipo de análise.

```
O Socrates tem estado horrível, mas o Louçã está em grande!
```

Listagem 3: Frase de exemplo.

```
<F ID="1">
   <VALUE>
    O Socrates tem estado horrível, mas o Louçã
    está em grande!
  </VALUE>
  <TOKENS>
    <TOKEN TYPE="DA">O</TOKEN>
    <TOKEN TYPE="PNM">Socrates</TOKEN>
    <TOKEN TYPE="VAUX">tem</TOKEN>
    <TOKEN TYPE="PPT">estado</TOKEN>
    <TOKEN TYPE="ADJ">horrível</TOKEN>
    <TOKEN TYPE="PNT">,</TOKEN>
    <TOKEN TYPE="CJ">mas</TOKEN>
    <TOKEN TYPE="DA">o</TOKEN>
    <TOKEN TYPE="PNM">Louçã</TOKEN>
    <TOKEN TYPE="V">está</TOKEN>
    <TOKEN TYPE="PREP">em</TOKEN>
    <TOKEN TYPE="ADJ">grande</TOKEN>
    <TOKEN TYPE="PNT">!</TOKEN>
  </TOKENS>
</F>
```

Listagem 4: Resultado do processamento realizado pelo módulo de AM v1.0.

A listagem 3 representa um exemplo de uma frase a processar. O resultado do processamento desta pela versão 1.0 do módulo encontra-se na listagem 4. Para facilitar a compreensão torna-se necessária uma breve descrição da estrutura resultante.

`TOKENS` Lista de elementos que compõe a frase.

`TOKEN TYPE="xx"` O elemento `TOKEN` contém um termo da frase. Este pode conter uma palavra ou um sinal de pontuação. O atributo `TYPE` neste elemento indica qual a classe morfológica da palavra de acordo com o analisador sintático utilizado.

`F ID="xx"` As frases são delimitadas pelas etiquetas `<F>` e `</F>`. O `ID` da frase identifica esta univocamente.

`VALUE` Este elemento contém o texto original da frase.

AM v2.0

A segunda versão do módulo fez uso de uma ferramenta um pouco mais complexa, o PALAVRAS, descrito no ponto 2.2.2. Utilizando as capacidades desta ferramenta foi possível realizar a análise morfológica e sintática dos textos.

```
<F ID="1">
  <VALUE>
    O Socrates tem estado horrível, mas o Louçã
    está em grande!
  </VALUE>
  <CLAUSES>
    <CLAUSE TYPE="FCL">
      <GROUP_FORM TYPE="NP">
        <TOKEN TYPE="ART">O</TOKEN>
        <TOKEN TYPE="PROP">Socrates</TOKEN>
      </GROUP_FORM>
      <TOKEN TYPE="V-FIN" INFINITIVE="achar"
          CONJUGATION="0/1/3S">tem</TOKEN>
      <TOKEN TYPE="V-PCP">estado</TOKEN>
      <TOKEN TYPE="ADJ">horrível</TOKEN>
    </CLAUSE>
    <CLAUSE TYPE="FCL">
      <GROUP_FORM TYPE="NP">
        <TOKEN TYPE="ART">o</TOKEN>
        <TOKEN TYPE="PROP">Louçã</TOKEN>
      </GROUP_FORM>
      <TOKEN TYPE="V-FIN" INFINITIVE="estar"
          CONJUGATION="PR 3S IND">está</TOKEN>
      <TOKEN TYPE="PRP">em</TOKEN>
      <TOKEN TYPE="ADJ">grande</TOKEN>
    </CLAUSE>
  </CLAUSES>
</F>
```

Listagem 5: Resultado do processamento realizado pelo módulo de AM v2.0.

A listagem 5 mostra o resultado do processamento da frase contida na listagem 3 por esta versão do módulo. Para facilitar a compreensão são descritos os novos elementos relativamente à versão anterior.

`CLAUSES` O elemento `CLAUSES` contém uma lista das orações encontradas na frase.

`CLAUSE TYPE="FCL"` O elemento `CLAUSE` corresponde a uma oração da frase, composta por

sintagmas e/ou *tokens*. O atributo `TYPE` identifica o tipo da cláusula, que apenas pode tomar o valor de `FCL` que corresponde à oração finita. Existem outros tipos de oração que não foram tidos em conta na implementação do protótipo.

`GROUP_FORM TYPE="NP"` A etiqueta `GROUP_FORM` representa um sintagma. O atributo `TYPE` apenas pode tomar o valor de `NP`. Existem outros sintagmas que não foram tidos em conta na implementação, foi no entanto deixada em aberto a possibilidade da sua inclusão no futuro.

4.3 Reconhecedor de Entidades Mencionadas

Neste ponto descreve-se uma das componentes principais do sistema desenvolvido, o módulo de REM. Adicionalmente é mostrado o seu propósito e justificada a sua utilização.

4.3.1 Finalidade

Este módulo desempenha um papel de elevada importância na AS, permitindo identificar o objeto alvo do sentimento. Existem classificadores bastante atuais, principalmente os orientados para *microblogs*, que escolhem não abordar a temática da deteção de entidades [5, 20]. Ter em conta o REM, mesmo em fontes fracamente estruturadas, traz valor acrescentado à classificação, permitindo inclusive corrigir erros de contextualização [26].

Vamos considerar como exemplo a frase contida na listagem 6. Analisando a frase sem muito detalhe, é fácil perceber que existe uma entidade (*Cristiano Ronaldo*) e que o sentimento para com esta entidade é claramente negativo (*Odeio*). No entanto, suponhamos que não era efetuado o reconhecimento da entidade, e que o autor da frase coloca esta no contexto de um jornal desportivo. Facilmente o classificador reconhece a frase como negativa, no entanto irá associar incorretamente este sentimento ao jornal ao invés da entidade presente na frase.

```
1 Odeio o Cristiano Ronaldo!
```

Listagem 6: Exemplo de *tweet*.

Efetivamente, a identificação de entidades em algo como os *tweets* não é trivial, no entanto é considerável o valor acrescentado ao processo de classificação.

4.3.2 Recursos

Durante o desenvolvimento da componente de REM foi criado um recurso adicional digno de nota. Neste ponto é feita a descrição desse recurso e qual o propósito da sua utilização.

Catálogo de Entidades

O Catálogo de Entidades foi um recurso utilizado pelo protótipo no módulo de REM. A sua finalidade foi simplesmente agrupar um conjunto de formas distintas de mencionar determinada entidade. O recurso foi gerado manualmente através de uma observação das menções já identificadas no corpus em estudo.

Evidentemente, este recurso tem uma dependência elevada de domínio, serve no entanto como prova de conceito para o principal corpus analisado. A utilização de um recurso deste tipo de uma forma mais abrangente implicaria um maior nível de completude. Seria necessário incluir mais entidades de tipos distintos, informação temporal que assegurasse a validade da referência em determinado momento, e claro, um maior número de menções para as entidades. A inspiração para esta aproximação surgiu principalmente do recurso POWER-PT [32]. Este consiste numa ontologia que reúne informação sobre políticos, organizações políticas e eleições mencionados nos meios de comunicação. A simplicidade do recurso aqui apresentado não permite que lhe seja atribuída a categoria de ontologia, no entanto, a finalidade pretendida é algo semelhante: Reconhecer uma entidade através de uma menção, seja esta uma alcunha ou apenas parte de um nome.

```
1   <Entity NAME="Manuela Ferreira Leite">
2      <ALVO TIPO="NOME">Manuela Ferreira Leite</ALVO>
3      <ALVO TIPO="CARGO">Drª Ferreira Leite</ALVO>
4      <ALVO TIPO="NOME">MFleite</ALVO>
5      <ALVO TIPO="NOME">Manuela FL</ALVO>
6      <ALVO TIPO="NOME">Ferreira Leite</ALVO>
7      <ALVO TIPO="NOME">FL</ALVO>
8      <ALVO TIPO="NOME">MFL</ALVO>
9      <ALVO TIPO="CARGO">Drª Manuela</ALVO>
10     <ALVO TIPO="NOME">Ministra da Educação</ALVO>
11     <ALVO TIPO="CARGO">Ministra das Finanças</ALVO>
12  </Entity>
```

Listagem 7: Exemplo extraído do *SentiCorpus-PT-Entities*.

A listagem 7 ilustra a informação relativamente à entidade "*Manuela Ferreira Leite*", descrita de seguida.

`Entity NAME="xx"` Cada entidade tem a sua informação contida no elemento `Entity`, identificado pelas etiquetas `<Entity>` e `</Entity>`. O atributo `NAME` tem o valor correspondente ao nome real da entidade.

`ALVO` O elemento `ALVO`, identificado pelas etiquetas `<ALVO>` e `</ALVO>` tem como valor uma possível menção a entidade.

`TIPO="xx"` O atributo `TIPO` do elemento `ALVO`, identifica o tipo de menção. Estas podem ser `NOME`, `CARGO`, `ORG`, `ALCUNHA`, `PRON` ou `GN_livre`. Os vários tipos são os mesmos

definidos para o SentiCorpus-PT. Este será descrito no ponto 5.1.1.

4.3.3 Funcionamento

O funcionamento do módulo pode ser descrito em três passos simples:

1. Receber informação processada pelo analisador morfossintático;
2. Identificar menções a entidades;
3. Retornar entidade identificada e respetivas menções de forma estruturada.

O processamento realizado procura adicionar o máximo de informação útil ao texto original, de forma a facilitar a classificação posterior realizada pelo módulo de AS. A quantidade e a forma como esta informação é adicionada evolui ao longo das várias versões desenvolvidas, como descrito no ponto 4.3.4.

O funcionamento geral do módulo é ilustrado pela figura 4.3.

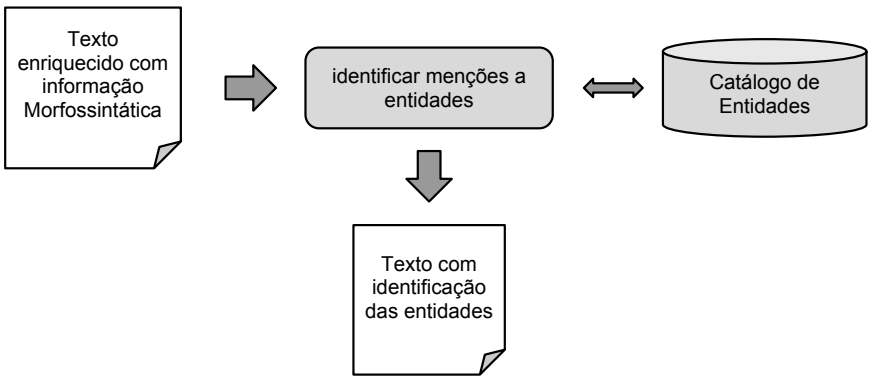

Figura 4.3: Funcionamento do módulo de REM.

4.3.4 Evolução

O desenvolvimento da componente de REM passou por várias etapas, fruto de uma sequência de aproximações com o objetivo de melhorar os resultados. Neste ponto são descritas e justificadas as escolhas efetuadas e o modo como foram implementadas em cada versão.

REM v1.0

A primeira versão deste módulo foi implementada tendo como principal objetivo a simplicidade. Esta abordagem permitiu visualizar os primeiros resultados e perceber, através

da sua análise, a direção a tomar nas versões seguintes.

Esta primeira versão do módulo fez já uso do recurso criado para auxiliar o processo, o *SentiCorpus-PT-Entities*[2] descrito no ponto 4.3.2. O método compreende dois passos.

1. Para cada entidade do *SentiCorpus-PT-Entities*, identificam-se ocorrências das respetivas menções no texto.
2. Para cada entidade encontrada, a frase é duplicada com informação relativa a entidade e ao modo como é mencionada.

```
<F ID="1" ALVO="José Sócrates">
  <ALVOS><ALVO TIPO="NOME">Socrates</ALVO></ALVOS>
  <VALUE>
    O Socrates tem estado horrível, mas o Louçã
    está em grande!
  </VALUE>
  <TOKENS>...</TOKENS>
</F>
<F ID="1" ALVO="Francisco Louçã">
  <ALVOS><ALVO TIPO="NOME">Louçã</ALVO></ALVOS>
  <VALUE>
    O Socrates tem estado horrível, mas o Louçã
    está em grande!
  </VALUE>
  <TOKENS>...</TOKENS>
</F>
```

Listagem 8: Frase exemplo processada pelo módulo de REM v1.0.

A listagem 8 mostra o resultado do processamento da frase contida na listagem 4 pelo módulo de REM v1.0. Foi omitida alguma informação irrelevante para este ponto com o objetivo de não tornar a listagem demasiado extensa. Descrevem-se os novos elementos resultantes do processamento.

`ALVO="José Sócrates"` O atributo `ALVO` da frase (`F`) indica qual a entidade identificada na frase.

`ALVOS` O elemento `ALVOS` contém a lista de menções encontradas que levaram à identificação da entidade na frase.

`ALVO TIPO="NOME"` O elemento `ALVO` identifica uma menção encontrada na frase. Indica adicionalmente o tipo de menção através do atributo `TIPO`. Este assume valores de acordo com o definido no catálogo de entidades.

[2] Versão específica do Catálogo de Entidades utilizada para o domínio analisado nos testes.

REM v2.0

A motivação para esta versão surge duma primeira tentativa de explorar as regras de construção frásica para melhorar os resultados. Nesta versão foi dado o primeiro passo para tentar tirar partido de duas componentes sintáticas: as orações[3], e o sintagma nominal[4].

Através da utilização destas estruturas gramaticais tentou-se não só melhorar o REM, como também passar a primeira noção de contexto para o módulo de AS. A título de exemplo na listagem 9 são identificadas duas orações finitas[5].

```
1     O Socrates tem estado horrível , mas o Louçã está em grande !
2    |__________|                          |_______|
3        sn1                                 sn2
4    |_______________________________|     |_______________________|
5                   of1                             of2
```

Listagem 9: Frase exemplo processada pelo módulo de REM v1.0.

Em ambas as orações, `of1` e `of2`, é possível encontrar sintagmas nominais, `sn1` e `sn2`, que contêm referência a uma entidade cada. Este facto é efetivamente benéfico para o que este módulo se propõe fazer. Adicionalmente, dentro da oração verifica-se a existência de verbos, "*ter*" e "*estar*", que fazem a ligação entre os sintagmas nominais e os adjetivos, "*horrível*" e "*grande*". Esta informação é posteriormente tida em consideração pelo módulo de AS.

A listagem 10 mostra o resultado do processamento realizado por esta versão.

REM v3.0

A versão 3.0 surge da necessidade de melhorar os resultados da versão 2.0. Ao contrário do previsto, a versão 2.0 não obteve o melhor desempenho. As razões do sucedido são analisadas no capítulo 5.

Nesta versão, mantendo um pouco a direção da anterior, o foco continua a ser a estrutura sintática das frases. A alteração passa por deixar de fora o sintagma nominal e utilizar apenas as orações. Esta aproximação é bastante mais abrangente, permitindo encontrar entidades fora dos sintagmas nominais, que de outra forma seriam descartadas. Ainda assim, continua a permitir relacionar de forma direta a entidade com uma possível polaridade. Tendo em conta que a análise será feita ao nível da oração, a probabilidade de um sentimento identificado ser direcionado à entidade identificada espera-se alta.

[3]Uma oração consiste em todo o conjunto linguístico que se estrutura em torno de um verbo, contendo sujeito e predicado.

[4]O sintagma nominal é um grupo de palavras cujo núcleo é constituído por um substantivo ou pronome.

[5]Uma oração finita é uma oração que contém informação de tempo, podendo ocorrer isolada, como oração principal ou subordinada. Ex: O João é simpático.

```
<F ID="1" ALVO="José Sócrates">
  <ALVOS><ALVO TIPO="NOME">Socrates</ALVO></ALVOS>
  <VALUE>
    O Socrates tem estado horrível, mas o Louçã
    está em grande!
  </VALUE>
  <CLAUSES>
    <CLAUSE TYPE="FCL">
      <GROUP_FORM TYPE="NP" ALVO="José Sócrates">
        <TOKEN TYPE="ART">O</TOKEN>
        <TOKEN TYPE="PROP">Socrates</TOKEN>
      </GROUP_FORM>
      <TOKEN TYPE="V-FIN">tem</TOKEN>
      <TOKEN TYPE="V-PCP">estado</TOKEN>
      <TOKEN TYPE="ADJ">horrível</TOKEN>
    </CLAUSE>
  </CLAUSES>
</F>
<F ID="1" ALVO="Francisco Louçã">
  <ALVOS><ALVO TIPO="NOME">Louçã</ALVO></ALVOS>
  <VALUE>
    O Socrates tem estado horrível, mas o Louçã
    está em grande!
  </VALUE>
  <CLAUSES>
    <CLAUSE TYPE="FCL">
      <GROUP_FORM TYPE="NP" ALVO="Francisco Louçã">
        <TOKEN TYPE="ART">o</TOKEN>
        <TOKEN TYPE="PROP">Louçã</TOKEN>
      </GROUP_FORM>
      <TOKEN TYPE="V-FIN">está</TOKEN>
      <TOKEN TYPE="PRP">em</TOKEN>
      <TOKEN TYPE="ADJ">grande</TOKEN>
    </CLAUSE>
  </CLAUSES>
</F>
```

Listagem 10: Frase exemplo processada pelo módulo de REM v2.0.

A listagem 11 ilustra o resultado do processamento da frase com esta versão do módulo de REM.

REM v4.0

Na última versão implementada para este módulo procurou-se melhorar os resultados indo de encontro às funcionalidades do analisador sintático utilizado, o PALAVRAS, já

```
<F ID="1" ALVO="José Sócrates">
  <ALVOS><ALVO TIPO="NOME">Socrates</ALVO></ALVOS>
  <VALUE>
    O Socrates tem estado horrível, mas o Louçã
    está em grande!
  </VALUE>
  <CLAUSES>
    <CLAUSE TYPE="FCL">
      <TOKEN TYPE="ART">O</TOKEN>
      <TOKEN TYPE="PROP">Socrates</TOKEN>
      <TOKEN TYPE="V-FIN">tem</TOKEN>
      <TOKEN TYPE="V-PCP">estado</TOKEN>
      <TOKEN TYPE="ADJ">horrível</TOKEN>
    </CLAUSE>
  </CLAUSES>
</F>
<F ID="1" ALVO="Francisco Louçã">
  <ALVOS><ALVO TIPO="NOME">Louçã</ALVO></ALVOS>
  <VALUE>
    O Socrates tem estado horrível, mas o Louçã
    está em grande!
  </VALUE>
  <CLAUSES>
    <CLAUSE TYPE="FCL">
      <TOKEN TYPE="ART">o</TOKEN>
      <TOKEN TYPE="PROP">Louçã</TOKEN>
      <TOKEN TYPE="V-FIN">está</TOKEN>
      <TOKEN TYPE="PRP">em</TOKEN>
      <TOKEN TYPE="ADJ">grande</TOKEN>
    </CLAUSE>
  </CLAUSES>
</F>
```

Listagem 11: Frase exemplo processada pelo módulo de REM v3.0.

mencionado no ponto 2.2.2.

Foi realizado um pré-processamento dos tweets, substituindo as menções a entidades identificadas pelo seu nome completo. Este passo inicial teve como objetivo facilitar a análise morfológica, permitindo um reconhecimento dos nomes próprios mais direto por parte do PALAVRAS. Através deste método esperou-se recolher orações até então não identificadas.

A listagem 12 mostra o resultado do pré-processamento efetuado. O valor do elemento `VALUE-NER-REPLACED` é desta feita passado ao analisador sintático ao invés da frase original como anteriormente.

```
<F ID="1">
  <VALUE>
    O Socrates tem estado horrível, mas o Louçã
    está em grande!
  </VALUE>
  <VALUE-NER-REPLACED>
    O José Socrates tem estado horrível, mas o Francisco
    Louçã está em grande!
  </VALUE-NER-REPLACED>
</F>
```

Listagem 12: Frase de exemplo, depois de aplicado o pré-processamento do REM v4.0.

4.4 Analisador de Sentimento

A componente de AS é efetivamente a componente mais importante de todo o modelo, não retirando o protagonismo das restantes, é aqui que se define qual o resultado final de todo o sistema.

4.4.1 Finalidade

A finalidade deste módulo consiste na identificação do tipo de opinião presente relativamente à entidade mencionada previamente encontrada. Para atingir o seu objetivo é necessário que sejam satisfeitas duas premissas fundamentais:

- **Existência de alvo**

  Apenas são classificadas expressões nas quais foram identificados alvos. Este componente utiliza assim o trabalho realizado pelo módulo de REM, procurando associar às entidades mencionadas o sentimento respetivo.

- **Subjetividade**

  Expressões objectivas encontram-se regularmente isentas de sentimento ou opinião, este módulo procura detetar e classificar apenas conteúdo subjetivo.

Caso estes requisitos sejam cumpridos, procura-se atribuir uma de três classes ao texto em análise: *positivo*, *negativo* ou *neutro*.

4.4.2 Recursos

Para criar a componente de AS foram utilizados alguns recursos adicionais. Nesta secção são descritos esses mesmos recursos no formato em que são disponibilizados, sendo também descrito o propósito da sua utilização.

SentiLex-PT

O SentiLex-PT[6] [43] é um léxico de sentimentos desenvolvido no âmbito do projeto REACTION[7]. Este já foi referido anteriormente, a título de exemplo no ponto 2.1.4, e foi o escolhido para fazer parte do módulo de AS do protótipo desenvolvido.

O SentiLex-PT é composto por 7.014 lemas e 82.347 formas flexionadas. De forma mais concreta é consiste em:

- 4.779 (16.863) adjetivos;
- 1.081 (1.280) nomes;
- 489 (29.504) verbos;
- 666 (34.700) expressões idiomáticas.

As entradas que compõem o léxico foram reunidas a partir de vários léxicos e corpora[8] disponíveis publicamente.

Para cada entrada os atributos relevantes são:

- O alvo do sentimento;
- A polaridade atribuída;
- O método de atribuição da polaridade.

A informação de polaridade foi maioritariamente atribuída de forma manual, no entanto alguns adjetivos tiveram a sua polaridade atribuída por uma ferramenta desenvolvida para o efeito.

O léxico é disponibilizado em dois ficheiros no formato TXT[9], o *SentiLex-lem-PT02* e o *SentiLex-flex-PT02*. Exemplos do conteúdo de ambos os documentos são incluídos nas listagens 13 e 14.

SentiLex-lem-PT02 Contém informação de polaridade sobre os lemas, sendo que cada linha inclui:

- Lema: normalmente a forma masculina do singular para os adjetivos, a forma singular para os nomes que flexionam em número e a forma infinitiva para os verbos e expressões idiomáticas.

[6] Informação disponível em http://dmir.inesc-id.pt/project/SentiLex-PT_02_in_English.

[7] *Retrieval, Extraction and Aggregation Computing Technology for Integrating and Organizing News.* Informação disponível em http://dmir.inesc-id.pt/project/Reaction.

[8] Plural de "corpus".

[9] Ficheiro de texto.

- Categoria gramatical: indicado através da chave `PoS` podendo ser-lhe atribuído o valor de adjetivo (`ADJ`), nome (`N`), verbo (`V`) ou expressão idiomática (`IDIOM`).
- Atributos de sentimento:
  - Polaridade: designada por `POL` e podendo ser positiva, negativa ou neutra e assumindo os valores numéricos `1`, `-1` e `0`, respetivamente.
  - Alvo da polaridade: designado por `TG` a qual corresponde um nome de tipo humano a que se atribui o valor de `HUM`. Adicionalmente são indicadas as possíveis funções do alvo, função de sujeito (`N0`) e/ou complemento (`N1`).
  - Classificação de polaridade: referenciada pela chave `ANOT`, indica se a polaridade foi atribuída de forma manual (`MAN`) ou de forma automática (`JALC`).

SentiLex-flex-PT02 Para cada entrada flexionada inclui a informação relativa ao lema presente no *SentiLex-lem-PT02*. Adicionalmente, quando a categoria gramatical é adjetivo, é dada informação sobre o género (masculino (`m`) ou feminino (`f`)) e número (singular (`s`) ou plural(`p`)). No caso em que a categoria gramatical é verbo, é incluída informação sobre o tempo, pessoa e número.

```
1  aberração.PoS=N;TG=HUM:N0;POL:N0=-1;ANOT=MAN
2  bonito.PoS=Adj;TG=HUM:N0;POL:N0=1;ANOT=MAN
3  castigado;PoS=Adj;TG=HUM:N0;POL:N0=-1;ANOT=JALC
4  estimado.PoS=Adj;TG=HUM:N0;POL:N0=1;ANOT=JALC;REV=AMB
5  enganar.PoS=V;TG=HUM:N0:N1;POL:N0=-1;POL:N1=0;ANOT=MAN
6  engolir em seco.PoS=IDIOM;TG=HUM:N0;POL:N0=-1;ANOT=MAN
```

Listagem 13: Exemplo extraído do *SentiLex-lem-PT02*.

```
1  aberração,aberração.PoS=N;FLEX=fs;TG=HUM:N0;POL:N0=-1;ANOT=MAN
2  bonita,bonito.PoS=Adj;FLEX=fs;TG=HUM:N0;POL:N0=1;ANOT=MAN
3  bonitas,bonito.PoS=Adj;FLEX=fp;TG=HUM:N0;POL:N0=1;ANOT=MAN
4  bonito,bonito.PoS=Adj;FLEX=ms;TG=HUM:N0;POL:N0=1;ANOT=MAN
5  bonitos,bonito.PoS=Adj;FLEX=mp;TG=HUM:N0;POL:N0=1;ANOT=MAN
```

Listagem 14: Exemplo extraído do *SentiLex-flex-PT02*.

No momento da escrita desta dissertação o léxico encontra-se na sua versão 2 e está disponível mediante pedido.

O SentiLex-PT foi utilizado pelo módulo de AS desde o início do seu desenvolvimento, fazendo parte integrante da estratégia escolhida para o funcionamento da componente. A sua função consistiu em permitir, de uma forma simples, identificar a polaridade de uma palavra isoladamente.

Admite-se a existência de algumas limitações na utilização deste recurso. Aquela que é possivelmente a mais óbvia, é a inexistência de um grau de confiança. O valor absoluto da polaridade poderia ser substituído por um valor entre 0 e 1 para cada umas das três classes

utilizadas. Esta aproximação é utilizada por outros léxicos e recursos semelhantes, como por exemplo o SentiWordNet [17], já mencionado anteriormente no ponto 2.1.4. Uma outra limitação percetível aquando da sua utilização é o reduzido número de entradas que compõe, sendo bastante fácil encontrar elementos comuns para os quais não existe classificação.

Dicionário de Sinónimos

Para determinar se dois termos são ou não sinónimos, o sistema recorre a uma rede semântica [38, 33] externa, consultada através de um serviço REST[10] que os autores disponibilizaram para este trabalho. Esta rede semântica possui aproximadamente 200.000 relações de sinonímia, para além de outras informações semânticas, como hiperonímia. Com a utilização deste recurso procurou-se classificar elementos que não encontrados no léxico de sentimentos, minimizando uma das suas fragilidades.

4.4.3 Funcionamento

O módulo foi desenhado tendo como um dos objetivos a simplicidade, procurando amenizar o trabalho necessário em futuras otimizações.

O módulo é alimentado com informação processada pela componente de REM, desta forma garantindo a existência de alvos para a opinião a ser agora encontrada. Dependendo da versão do módulo de REM o formato varia ligeiramente, fator compreensível tendo em conta que o enriquecimento aumenta à medida que o módulo evoluiu. De uma forma geral a informação recebida é composta por 3 elementos fundamentais:

- Alvo identificado;
- Contexto que o alvo é mencionado;
- Classificação morfológica dos elementos do contexto.

O contexto em que o alvo é mencionado é analisado, identificando elementos candidatos a passar pelo processo de polarização. Para realizar a escolha destes elementos é utilizada a sua classificação morfológica previamente realizada. Os critérios utilizados nesta decisão foram alterados ao longo das várias versões desta componente, as alterações são descritas ao longo do ponto 4.4.4.

Depois de apurados os elementos a polarizar é iniciado o processo de atribuição da classe a cada um destes. Como referido anteriormente, são utilizadas três classes: *positivo*, *negativo* e *neutro*. Numericamente, cada uma destas classes é identificada pelo valor 1,

[10] Do acrónimo *Representational State Transfer*, refere-se a uma arquitetura normalmente utilizada em sistemas distribuídos.

-1 e 0, respetivamente. Para atribuir o valor numérico a cada elemento é utilizado um dos recursos já referidos, o SentiLex-PT. A cada elemento classificável é atribuído o valor definido no léxico de sentimento, preparando a fase final do processo. Devido ao facto do léxico não possuir classificação para todos os elementos definidos como polarizáveis pelo módulo, tornou-se necessária a utilização de um outro recurso adicional, um dicionário de sinónimos. Este é descrito no ponto 4.4.2 e utilizado para colmatar esta fragilidade do SentiLex-PT.

Definidos os valores numéricos da polaridade para cada elemento polarizável, é calculado o valor global para polaridade do contexto. Este valor determina qual a classe a atribuir ao conteúdo e, consequentemente, o tipo de opinião que incide sobre a entidade identificada. Na eventualidade de não ter sido possível identificar a polaridade de um elemento, atribui-se a classificação de *neutro* ao mesmo. A fórmula de cálculo sofreu alterações ao longo da evolução do módulo, sendo que algumas tiveram um impacto fundamental na performance de todo o sistema. Os vários métodos de cálculo são descritos na secção 4.4.4 para as várias versões do sistema.

O funcionamento desta componente é graficamente ilustrado pela figura 4.4.

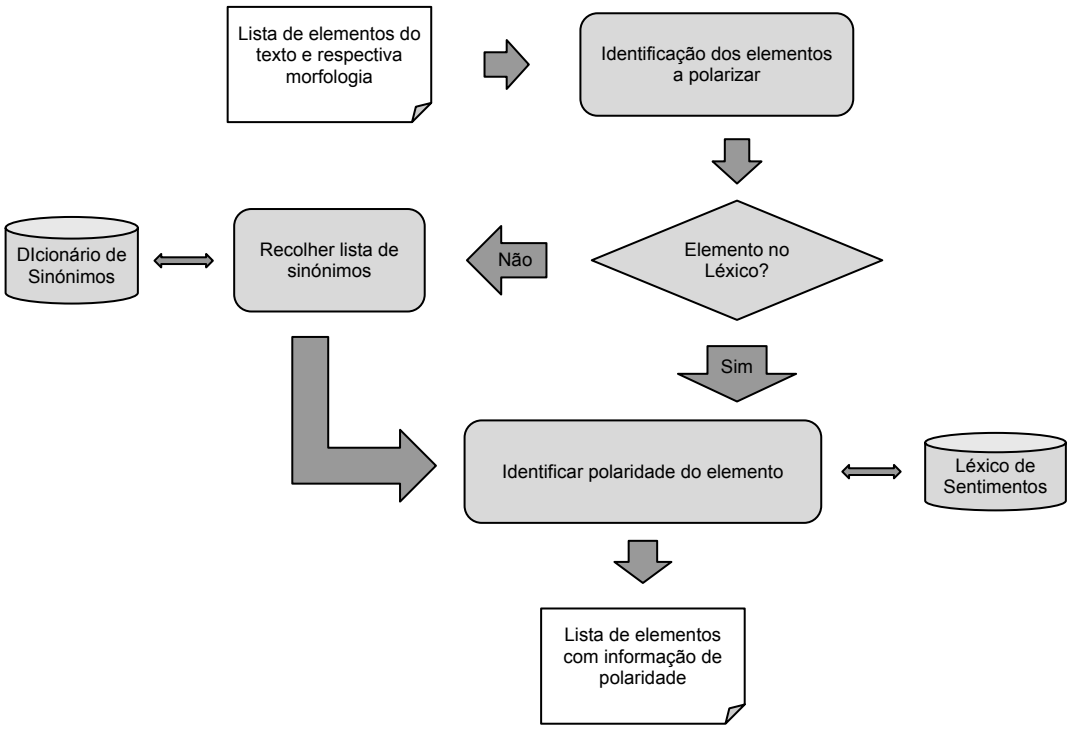

Figura 4.4: Funcionamento do módulo de AS.

4.4.4 Evolução

Tendo em conta tratar-se da componente que efetivamente estabelece qual veredicto no que diz respeito à polaridade do *input* recebido, mereceu o maior foco em desenvolvimento e investigação. Nesta secção expressam-se as várias iterações evolutivas pelas quais passou o protótipo, procurando demonstrar as razões e motivações que definiram o rumo tomado.

AS v1.0

A primeira versão do módulo de sentimento foi também a sua aproximação mais simples. Procurou-se utilizar a componente morfológica da língua que mais naturalmente se associa a expressão de opinião, o adjetivo. Não foram consideradas quaisquer propriedades sintáticas do texto e o conteúdo fornecido foi abordado como atómico. Este fator criou problemas logo a partida quando na mesma frase se encontravam várias entidades. A listagem 15 contém um exemplo de uma frase que alimenta esta versão da componente.

```
<F ID="1" ALVO="Francisco Louçã">
  <ALVOS><ALVO TIPO="NOME">Louçã</ALVO></ALVOS>
  <VALUE>
    O Socrates tem estado horrível, mas o Louçã
    está em grande!
  </VALUE>
  <TOKENS>
      <TOKEN TYPE="ADJ">horrível</TOKEN>
      <TOKEN TYPE="ADJ">grande</TOKEN>
  </TOKENS>
</F>
```

Listagem 15: Frase de exemplo, versão 1.0 do módulo de AS.

Como já mencionado, a aproximação escolhida nesta versão leva a que caso se verifique a existência de várias entidades na mesma frase, a probabilidade de acerto seja mínima. Na frase contida na listagem 15 existem efetivamente duas entidades mas será impossível associar a qual destas se refere o adjetivo.

Como resultado desta aproximação o resultado da classificação da frase da listagem 15 será o apresentado na listagem 16. A polaridade "0", ou seja classificação de *neutro*, está incorreta resultando diretamente da inabilidade de associar um contexto em torno da entidade por parte da versão atual do módulo.

AS v2.0

No seguimento das falhas detetadas pela versão anterior, nomeadamente a incapacidade de identificar o alvo a que se aplica o elemento de opinião, esta iteração procurou solucionar

```
<F ID="1" ALVO="Francisco Louçã" POL="0">
  <ALVOS><ALVO TIPO="NOME">Louçã</ALVO></ALVOS>
  <VALUE>
    O Socrates tem estado horrível, mas o Louçã
    está em grande!
  </VALUE>
  <TOKENS>
      <TOKEN TYPE="ADJ" POL="-1">horrível</TOKEN>
      <TOKEN TYPE="ADJ" POL="1">grande</TOKEN>
  </TOKENS>
</F>
```

Listagem 16: Frase de exemplo, versão 1.0 do módulo de AS.

o problema recorrendo às funções sintáticas da língua.

Utilizando a informação já recolhida no módulo de REM na sua versão 2.0, descrita no ponto 4.3.4, foi considerado como contexto do alvo a oração em que este se insere.

```
<F ID="1" ALVO="Francisco Louçã">
  <ALVOS><ALVO TIPO="NOME">Louçã</ALVO></ALVOS>
  <VALUE>
    O Socrates tem estado horrível, mas o Louçã
    está em grande!
  </VALUE>
   <CLAUSES>
    <CLAUSE TYPE="FCL">
      <TOKEN TYPE="ADJ">grande</TOKEN>
    </CLAUSE>
  </CLAUSES>
</F>
```

Listagem 17: Frase de exemplo antes de processada, versão 2.0 do módulo de AS.

Como é visível na listagem 17, esta aproximação permitiu associar ao alvo o adjetivo que o influência.

Como resultado do isolamento do contexto do alvo, o exemplo da listagem 17 é corretamente classificado. O resultado é visível na listagem 18.

AS v3.0 e v4.0

As versões 3.0 e 4.0 da componente de AS resultaram da tentativa de classificação de frases que não contivessem adjetivos. Para atingir este fim procurou-se utilizar outras classes gramaticais das quais fosse possível extrair algum nível de opinião. As classes eleitas foram os nomes e os verbos.

```
<F ID="1" ALVO="Francisco Louçã" POL="1">
  <ALVOS><ALVO TIPO="NOME">Louçã</ALVO></ALVOS>
  <VALUE>
    O Socrates tem estado horrível, mas o Louçã
    está em grande!
  </VALUE>
   <CLAUSES>
    <CLAUSE TYPE="FCL">
      <TOKEN TYPE="ADJ" POL="1">grande</TOKEN>
    </CLAUSE>
  </CLAUSES>
</F>
```

Listagem 18: Frase de exemplo depois de processada, versão 2.0 do módulo de AS.

A capacidade de transmissão de opinião da classe gramatical dos verbos é já um dado adquirido, desta forma não é algo de novo a sua utilização na deteção de polaridade [28, 13]. A escolha da classe nome entendeu-se como natural, sendo do senso comum a existência de nomes com conotação negativa ou positiva.

A listagem 19 mostra alguns exemplos das classes escolhidas que fazem parte do SentiLex-PT.

```
aberração,aberração.PoS=N;FLEX=fs;TG=HUM:N0;POL:N0=-1;ANOT=MAN
adultério,adultério.PoS=N;FLEX=ms;TG=HUM:N0;POL:N0=-1;ANOT=MAN
afundas,afundar-se.PoS=V;Flex=P:2s;TG=HUM:N0;POL:N0=-1;ANOT=MAN
agredido,agredir.PoS=V;Flex=K;TG=HUM:N0:N1;POL:N0=-1;POL:N1=0;ANOT=MAN
```

Listagem 19: Exemplo extraído do *SentiLex-flex-PT02*.

AS v5.0

Com o objetivo de explorar mais propriedades linguísticas, na versão 5.0 procurou-se fazer uso da negação. A negação é uma construção da língua regularmente utilizada pela sua capacidade de alterar o sentido de uma oração. Esta característica torna-a bastante relevante na classificação de opiniões [51]. Por exemplo, é percetível que a frase 5.1 representa uma opinião positiva.

5.1 Eu $[gosto]^{+}$ de viajar.

O elemento "*gosto*", conjugação do verbo *gostar*, possui uma conotação positiva que é fácil de identificar. No entanto se for adicionado um elemento de negação o cenário inverte-se como se pode verificar pela frase 5.2.

5.2 Eu $[não\ gosto]^{-}$ de viajar.

A inserção do advérbio de negação "*não*" inverte a polaridade previamente identificada do verbo "*gosto*", tornando o sentimento do sujeito para com a objeto "*viajar*" negativo.

Seguindo o mesmo raciocínio da inclusão dos advérbios de negação, procurou-se utilizar outros elementos da língua que tenham a capacidade de alterar de alguma forma a intensidade da opinião presente na frase. Para o efeito foram utilizados alguns advérbios de intensidade. A frase 5.3 mostra um exemplo da aplicação deste tipo de modificadores de polaridade.

5.3 Viajar é $[bom]^{+}$ mas $[muito\ cansativo]^{--}$.

Aquando da deteção de um destes elementos, aplica-se uma alteração à polaridade dos elementos polarizáveis subsequentes dentro duma janela definida. Dependendo do tipo de advérbio a polaridade pode ser aumentada, diminuída ou invertida. De seguida enumeram-se os vários elementos considerados no módulo.

Advérbios de negação

- Não
- Nunca
- Negativamente

Locuções adverbiais de negação

- De modo algum
- De jeito nenhum
- De forma nenhuma

Advérbios de intensidade

- Bastante
- Demais
- Mais
- Menos
- Muito
- Quase
- Pouco
- Demasiado
- Imenso

Locuções adverbiais de intensidade ou quantidade

- Em excesso
- De todo
- De muito
- Por completo
- Por demais

O peso que o modificador aplica foi definido sem a aplicação de nenhum critério especial, pelo que se admite a necessidade de alguma investigação adicional para identificar quais os valores ótimos. A exceção poderá ser atribuída aos elementos pertencentes ao grupo dos advérbios de negação, sendo aceitável que estes realizem uma inversão de polaridade completa.

A janela até onde a influência do modificador se estende é algo que também merece algum estudo adicional. Nos testes efetuados foram testados vários valores, inclusive uma aproximação em que esta era dinâmica, mantendo-se até ao sinal de pontuação seguinte.

AS v6.0

A evolução para esta versão da componente não foi muito significativa, tratando-se apenas de uma constatação da influência das alterações efetuadas ao módulo de REM v4.0 descrito no ponto 4.3.4. As melhorias resultam da maior capacidade de satisfazer as condições necessárias para a classificação por parte do módulo, nomeadamente no que diz respeito a identificação de alvos.

AS v7.0

Nesta versão procurou-se efetuar uma distinção da importância atribuída às várias classes gramaticais tidas como polarizáveis até ao momento - nomes, verbos e adjetivos. A oração 7.1 mostra um dos exemplos incorretamente classificados que fundamentam esta aproximação.

7.1 Socrates $[errou]^{-}_{verb}$ o $[necessário]^{+}_{adj}$

Os elementos relevantes para a classificação nesta oração são "*errou*" - conjugação do verbo "*errar*" - e o adjetivo "*necessário*", sendo que estes possuem polaridades inversas de acordo com o SentiLex-PT. Considerando uma classificação em que a cada elemento é atribuída a mesma importância obtemos:

$$P_{7.1} = P(errou) + P(necessário) = 0$$

Em que $P_{7.1}$ representa a polaridade da oração 7.1 e *P(x)* representa a polaridade do elemento x, sendo que x toma o valore de "*errou*" e "*necessário*" neste exemplo. A versão 6.0 do classificador atribui a esta oração, incorretamente, uma polaridade com o valor numérico de 0, que corresponde à classe *neutro*. Face a este cenário procurou-se atribuir uma importância diferenciada aos elementos polarizáveis. Os valores numéricos inerentes a esta categorização foram alterados ao longo dos testes, não se considerando que tenham sido encontrados valores definitivos para o peso de cada classe. No entanto teve-se como correta a seguinte ordem decrescente de relevância das três classes gramaticais utilizadas:

1. Adjetivos
2. Nomes
3. Verbos

No seguimento desta alteração, a classificação atribuída a oração 7.1 pela presente versão do classificador assume um valor diferente.

$$P_{7.1} = P(errou) + P(necessário) \simeq P(necessário)$$

O que significa que o valor numérico que representa a polaridade da oração é agora semelhante ao da polaridade do adjetivo "*necessário*". Isto implica que a classe atribuída pela nova versão do classificador é *positivo*, o que está correto.

AS v8.0

A última versão implementada da componente de AS procurou amenizar as limitações do léxico de sentimentos utilizado, o SentiLex-PT. Para o efeito recorreu-se à utilização do recurso descrito no ponto 4.4.2. Este método permitiu polarizar elementos que não faziam parte do léxico, aumentando assim a capacidade de todo o sistema.

Para exemplificar o conceito vamos tomar como exemplo a oração 8.1.

8.1 Socrates esteve $[indiferente]^{-}_{adj}$

Da composição desta oração faz parte o adjetivo "*indiferente*" cuja polaridade no contexto é negativa. No entanto o SentiLex-PT não inclui este elemento, tornando assim impossível efetuar uma classificação correta. Numa tentativa de contornar esta situação recorreu-se a relação de sinonímia entre as palavras. Através desta relação existe a possibilidade de serem encontrados novos elementos polarizados pelo léxico e que, pela definição desta relação semântica, mantenham uma semelhança mínima de significado. Da lista de sinónimos classificados é atribuída ao elemento original a polaridade em maioria.

- Desapaixonado^{-}
- Frio^{-}
- Indolente^{-}
- Apático^{-}

Neste exemplo, todos os sinónimos encontrados estão presentes no léxico de sentimentos, e todos estes se encontram classificados como negativos. Posto isto, ao adjetivo "*indiferente*" seria atribuída polaridade numérica de -1 e consequentemente atribuída de forma correta a classe *negativo* à oração.

Existe uma probabilidade de erro a considerar, no entanto a utilização deste método está diretamente relacionada com as limitações do léxico de sentimentos. O aumento de completude deste implicaria uma diminuição da necessidade de recorrer a esta aproximação, diminuindo as hipótese de erro. A exploração das relações semânticas entre palavras para construção e enriquecimento de léxicos tem sido alvo de alguns trabalhos recentes, mostrando a sua viabilidade [39, 44].

4.5 Conclusões

Em suma, a solução desenvolvida é composta por três módulos independentes no seu funcionamento, mas complementares nas suas funções. Isto significa que cada módulo pode ser executado de forma independente sem a necessidade de conhecer o estado dos restantes. A comunicação entre os componentes é realizada através da troca de mensagens no formato XML, assegurando a independência tecnológica dos interlocutores caso seja necessário.

Existem várias questões deixadas em aberto pela solução apresentada, consequências da complexidade da matéria abordada e do compromisso temporal assumido na realização desta dissertação. Algumas destas questões são abordadas no capítulo 6, abrindo as portas a possíveis desenvolvimentos futuros.

Capítulo 5

Avaliação

Neste capítulo descrevem-se as várias experiências efetuadas durante o processo de construção do protótipo que ilustra esta dissertação. São enumerados os resultados obtidos e analisados de acordo com o seu impacto no módulo em que foram introduzidos e no sistema final.

5.1 Corpus Utilizados

Para realizar o teste e avaliação do sistema desenvolvido foram utilizados dois corpus anotados.

5.1.1 SentiCorpus-PT

O SentiCorpus-PT[1] é um corpus desenvolvido no âmbito do projecto REACTION[2], constituído por comentários a notícias de foro político. Está totalmente em Português e encontra-se anotado com informação de sentimento e de entidades mencionadas.

O corpus resulta da compilação de comentários efetuados por leitores do jornal "Público", no âmbito de 10 artigos noticiosos que cobriram os debates políticos que antecederam as eleições legislativas de 2009, em Portugal. Os debates tiveram lugar entre o dia 2 e o dia 12 de Setembro de 2009, onde participaram os candidatos a primeiro-ministro dos cinco partidos portugueses com maior representatividade no Parlamento [10].

[1] Informação disponível em http://dmir.inesc-id.pt/project/SentiCorpus-PT_01_in_English.

[2] Retrieval, Extraction and Aggregation Computing Technology for Integrating and Organizing News. Informação disponível em http://dmir.inesc-id.pt/project/Reaction.

A coleção completa é composta por cerca de 8000 frases agrupadas em 2.795 comentários. Os comentários encontram-se associados aos respetivos artigos de notícia.

A anotação encontra-se num formato concebido especificamente para o efeito. A anotação de sentimento inclui várias dimensões, tais como a polaridade do sentimento, o alvo da opinião, a menção do alvo e a presença eventual de ironia verbal.

O corpus manualmente anotado é composto por 3.888 frases com opinião, estas foram retiradas de 1.082 comentários dos leitores.

A listagem 20 inclui um pequeno extrato constituído por um comentário com duas frases, adicionalmente inclui-se uma descrição do formato apresentado.

```
<COMENT ID="1398874::1">
  <F ID="1" ALVO="José Sócrates" POL="1" INT="Literal">
    Continuo a achar que temos <ALVO TIPO=\GN_livre">Homem</ALVO> e
    <ALVO TIPO="CARGO">1° Ministro</ALVO>.
  </F>
  <F ID="2" ALVO="Paulo Portas" POL="-1" INT="Literal">
    O <ALVO TIPO="ALCUNHA">Paulinho das Feiras</ALVO> ,  como
    sempre demagogo.
  </F>
</COMENT>
```

Listagem 20: Exemplo extraído do *SentiCorpus-PT*.

`COMMENT ID="xxx::xx"` Cada comentário está contido entre as etiquetas `<COMMENT>` e `</COMMENT>`. O ID de cada comentário consiste na identificação do artigo noticioso a que se refere e no identificador do utilizador. Estes são separados pelo símbolo "`::`".

`F ID="xx"` Os comentários englobam um conjunto de uma ou mais frases. As frases são delimitadas pelas etiquetas `<F>` e `</F>`. O `ID` da frase identifica esta dentro do comentário através da sua posição no mesmo.

`ALVO="José Sócrates"` Este é um atributo da frase (`F`) e indica qual a entidade mencionada. É relativamente a esta entidade que se manifesta a opinião identificada pela polaridade atribuída à frase. Uma frase com opinião pode conter vários alvos. Maioritariamente os alvos mencionados correspondem aos políticos que participaram nos debates televisivos, podem no entanto ser referenciadas outras entidades mediáticas relevantes para o contexto.

`POL="1"` O atributo `POL` define a polaridade da frase (`F`) relativamente ao alvo mencionado e identificado pelo atributo `ALVO`. A polaridade e intensidade da opinião é representada através de um valor numérico, que pode ir desde `-2` (o valor negativo mais intenso) até ao `2` (o valor positivo mais intenso). As opiniões neutras são classificadas com `0`.

`INT="Literal"` O atributo `INT` identifica qual a interpretação da frase (`F`), esta pode ser literal (`Literal`), ou irónica (`Irónico`), sempre que apresente um significado diferente daquele que deriva da interpretação literal do texto.

`ALVO TIPO="ALCUNHA"` O elemento alvo identifica a menção a uma entidade na frase (`F`). As categorias sintático-semânticas identificadas são:

- `NOME`: Nome próprio ou acrónimo, estes podem ser precedidos de um nome de título ou de cargo;
- `CARGO`: Um nome ou expressão que remeta para um cargo ou posição profissional;
- `ORG`: O nome de uma organização;
- `ALCUNHA`: Uma alcunha;
- `PRON`: Um pronome;
- `GN_livre`: Um grupo nominal livre cuja referência pode ser interpretada ao nível da frase ou do comentário, após resolução anafórica e de correferência.

Adaptação

Para adaptar o SentiCorpus-PT ao modelo pretendido e facilitar o processo de avaliação dos resultados, foram realizadas algumas adaptações ao formato original do corpus.

O mesmo excerto apresentado na listagem 20 referente ao ponto 5.1.1 assume agora a nova forma apresentada na listagem 21. A este novo formato deu-se o nome de *SentiCorpus-PT-gold*.

As principais alterações realizadas foram:

`CORPUS` Todo o corpus é delimitado pelo elemento `CORPUS` através das suas respetivas etiquetas, `<CORPUS>` e `</CORPUS>`.

`COMENT ID` O identificador dos comentários (`COMMENT`) foi alterado para uma numeração sequencial. Esta alteração facilita a identificação dos elementos enquanto continua a identificar univocamente um comentário.

`ALVOS` Foi incluído um novo elemento, `ALVOS`, este utiliza as etiquetas `<ALVOS>` e `</ALVOS>`. O seu propósito é agregar as menções a uma entidade ocorridas na frase (`F`).

`ALVO` O elemento `ALVO`, delimitado pelas etiquetas `<ALVOS>` e `</ALVOS>`, tem como propósito identificar uma menção a uma entidade ocorrida na frase (`F`).

`VALUE` O texto original da frase encontra-se rodeado pelas etiquetas `<VALUE>` e `</VALUE>` dentro do elemento da frase (`F`).

```
<CORPUS>
 <COMENT ID="1">
    <F ID="1" ALVO="José Sócrates" POL="1" INT="Literal">
       <ALVOS>
          <ALVO TIPO="GN_livre">Homem</ALVO>
          <ALVO TIPO="CARGO">1º Ministro</ALVO>
       </ALVOS>
       <VALUE>Continuo a achar que temos Homem e 1º Ministro.</VALUE>
    </F>
    <F ID="2" ALVO="Paulo Portas" POL="-1" INT="Literal">
       <ALVOS>
          <ALVO TIPO="ALCUNHA">Paulinho das Feiras</ALVO>
       </ALVOS>
       <VALUE>O Paulinho das Feiras, como sempre demagogo.</VALUE>
    </F>
 </COMENT>
</CORPUS>
```

Listagem 21: Exemplo extraído do *SentiCorpus-PT-gold*.

Foi também criado um recurso que derivou do *SentiCorpus-PT-gold*. Este consiste numa versão não classificada do corpus a que se deu o nome de *SentiCorpus-PT-clean*. Este recurso é ilustrado pela listagem 22.

```
<CORPUS>
 <COMENT ID="1">
    <F ID="1" POL="0">
       <VALUE>Continuo a achar que temos Homem e 1º Ministro.</VALUE>
    </F>
    <F ID="2" POL="0">
       <VALUE>O Paulinho das Feiras, como sempre demagogo.</VALUE>
    </F>
 </COMENT>
</CORPUS>
```

Listagem 22: Exemplo extraído do *SentiCorpus-PT-clean*.

O *SentiCorpus-PT-clean* foi utilizado como *input* para o protótipo do classificador desenvolvido. A avaliação da classificação foi efetuada recorrendo a comparação com o *SentiCorpus-PT-gold*.

5.1.2 SentiTuites-PT

O SentiTuites-PT[3] foi, a semelhança do SentiCorpus-PT, um recurso desenvolvido no âmbito do projecto REACTION[4]. O corpus é composto por 30.470 tweets colocadas por utilizadores portugueses durante campanha para as eleições legislativas portuguesas de 2011. A criação deste corpus teve como objetivo recolher opiniões sobre os lideres políticos dos principais partidos durante as eleições:

- Pedro Passos Coelho (ppcoelho)
- José Sócrates (jsocrates)
- Paulo Portas (pportas)
- Jerónimo Sousa (jsousa)
- Francisco Louçã (flouca)

Os tweets foram recolhidos entre 29 de Abril de 2011 e 3 de Junho de 2011. Um subconjunto destas foi manualmente anotado relativamente ao sentimento para com o candidato mencionado. Os restantes tweets foram classificados automaticamente recorrendo a regras lexicais e sintáticas, utilizando um léxico de sentimentos e pistas para identificação de ironia e sarcasmo. A escala de polaridade compreende os valores 1, -1 e 0 que correspondem a sentimento positivo, negativo e neutro, respetivamente.

O corpus é disponibilizado em três ficheiros no formato CSV[5], utilizando como separador o símbolo "|". As listagens 23, 24 e 25 ilustram o formato desde recurso.

SentiTuites-tweets-01 Conjunto de tweets sem anotação de polaridade.

```
1 2011-04-29\2011-04-30|64369891096010752|20884910|flouca|francisco louçã
2 2011-06-02\2011-06-03|76370335771009024|34091495|ppcoelho|passos coelho
```

Listagem 23: Exemplo extraído do SentiTuites-PT. Ficheiro SentiTuites-tweets-01.csv

- `2011-04-29\2011-04-30`: Data em que foi recolhida;
- `64369891096010752`: Identificador do tweet;
- `20884910`: Identificador do utilizador que submeteu;
- `flouca`: Alvo do sentimento;
- `francisco louçã`: Modo como é mencionado;

SentiTuites-sentiment-01 Conjunto de tweets com anotação de polaridade atribuída automaticamente.

[3]Informação disponível em http://dmir.inesc-id.pt/project/SentiCorpus-PT_01_in_English.

[4]Retrieval, Extraction and Aggregation Computing Technology for Integrating and Organizing News. Informação disponível em http://dmir.inesc-id.pt/project/Reaction.

[5]Do Inglês *Comma-Separated Values*, valores separados por virgulas.

```
2011-04-29\2011-04-30|64369891096010752|0
2011-04-29\2011-04-30|64063070917492736|-1
```

Listagem 24: Exemplo extraído do SentiTuites-PT. Ficheiro SentiTuites-sentiment-01.csv

- Data em que foi recolhida;
- Identificador do tweet;
- Polaridade atribuída automaticamente;

SentiTuites-goldstandard-01 Conjunto de tweets com anotação de polaridade atribuída manualmente.

```
2011-05-17\2011-05-18|70601086792245249|-1
2011-05-17\2011-05-18|70600777852403712|1
```

Listagem 25: Exemplo extraído do SentiTuites-PT. Ficheiro SentiTuites-goldstandard-01.csv

- Data em que foi recolhida;
- Identificador do tweet;
- Polaridade atribuída manualmente;

Adaptação

De forma a tornar possível a utilização do SentiTuites-PT foi necessário proceder a algumas alterações do seu formato original. Para além das alterações a nível de formato, foi também cruzada informação dos vários recursos que compõem o corpus.

Foi criado o recurso *SentiTuites-PT-gold* ilustrado pela listagem 26. O formato deste é semelhante ao formato do *SentiCorpus-PT-gold* previamente abordado no ponto 5.1.1 e ilustrado na Listagem 21.

Para obter o *SentiTuites-gold* foi necessário assumir alguns compromissos. Inicialmente foi necessário recolher o texto dos tweets que compunham o *SentiTuites-goldstandard-01*, sendo que alguns destes já não se encontravam disponíveis. Dos tweets recolhidos foram utilizados todos aqueles que se encontravam referenciadas no *SentiTuites-tweets-01*. Deste modo foi possível obter informação relativamente às entidades mencionadas. De todos os tweets nestas condições foi utilizado um subconjunto, com os casos que possuíam apenas uma entidade mencionada identificada. A razão para esta limitação deve-se ao facto de não existir referência a que excerto da frase se refere a cada entidade. Este fator dificultaria o processo de avaliação e como tal optou-se por limitar a priori.

Tal como sucedeu com o SentiCorpus-PT, foi também criado um recurso a partir do *SentiTuites-PT-gold*, o *SentiTuites-PT-clean* ilustrado na listagem 27.

```
<CORPUS>
 <COMENT ID="1">
  <F ID="1" ALVO="Francisco Louçã" POL="1">
    <VALUE>Pode dizer-se tudo sobre Louçã mas não foi ele que levou
    o país à bancarrota.</VALUE>
  </F>
  <F ID="2" ALVO="Francisco Louçã" POL="-1">
    <VALUE>Eu votaria em Louçã se só o ouvisse 20 segundos em cada 3
    minutos.</VALUE>
  </F>
 </COMENT>
</CORPUS>
```

Listagem 26: Exemplo extraído do *SentiTuites-PT-gold*.

```
<CORPUS>
 <COMENT ID="1">
  <F ID="1" POL="0">
    <VALUE>Pode dizer-se tudo sobre Louçã mas não foi ele que levou
    o país à bancarrota.</VALUE>
  </F>
  <F ID="2" POL="0">
    <VALUE>Eu votaria em Louçã se só o ouvisse 20 segundos em cada 3
    minutos.</VALUE>
  </F>
 </COMENT>
</CORPUS>
```

Listagem 27: Exemplo extraído do *SentiTuites-PT-clean*.

O *SentiTuites-PT-clean* foi fornecido como *input* ao protótipo do classificador desenvolvido. O resultado da classificação foi avaliado através da comparação com o *SentiTuites-PT-gold*.

5.2 Métricas

Para proceder a uma avaliação de desempenho do modelo implementado foi necessário selecionar algumas métricas para o efeito. De forma a facilitar a comparação com outros estudos, foram escolhidas as métricas consideradas mais comuns na avaliação de tarefas de classificação. Os resultados são assim apresentados recorrendo a Tabelas de Confusão e através do cálculo dos valores referentes a Precisão, Cobertura e Medida-F [6].

5.2.1 Tabelas de Confusão

A Tabela de Confusão é uma tabela 2 x 2, que permite ilustrar de uma forma simples os resultados de um algoritmo ou sistema numa classificação a duas classes. Neste cenário, com a existência de duas classes, a tabela de confusão pode também ser denominada de tabela de contingência. As duas classes podem ser abordadas como apenas uma, a classe e a sua negação. A utilização desta tabela permite não só quantificar elementos incorretamente classificados como também dar uma perspetiva de qual poderá ser a origem do erro.

	Correto	**Não Correto**
Selecionado	VP	FP
Não Selecionado	FN	VN

Tabela 5.1: Tabela de Confusão.

Uma Tabela de Confusão, como a ilustrada pela Tabela 5.1, permite-nos observar os valores relativos aos Verdadeiros Positivos (VP), Falsos Positivos (FP), Falsos Negativos (FN) e Verdadeiros Negativos (VN). Estes valores têm o seguinte significado:

Verdadeiros Positivos O número de elementos da classe analisada que foram corretamente identificados;

Falsos Positivos O número de elementos identificados incorretamente como pertencente à classe analisada;

Falsos Negativos O número de elementos não identificados como pertencentes à classe analisada mas que pertencem efetivamente a esta;

Verdadeiros Negativos O número de elementos que foram corretamente identificados como não pertencentes à classe analisada.

[6] Do Inglês *F-measure*.

Estes valores permitem avaliar superficialmente a precisão da classificação, podem adicionalmente servir de base para o cálculo de outras métricas qualitativas.

5.2.2 Precisão

A Precisão é uma das métricas que classifica a qualidade da classificação efetuada. Esta métrica representa o balanço entre a quantidade de elementos corretamente identificados e todos os elementos identificados como pertencentes a uma classe. O seu cálculo pode ser efetuado através dos valores que compõem a Tabela de Confusão.

$$precisão = \frac{VP}{VP + FP}$$

Ao avaliar um sistema apenas pela sua Precisão é possível ser-se iludido quanto ao seu desempenho. Um valor de 1.0 (100%) nesta métrica significa que todos os elementos classificados como pertencentes a uma classe pertencem efetivamente a esta. No entanto nada nos indica relativamente aos elementos que deveriam ter sido identificados e não o foram.

5.2.3 Cobertura

A Cobertura é uma métrica que permite, tal como a Precisão, avaliar a qualidade de um classificador. O objetivo da Cobertura é quantificar os elementos que deveriam ter sido identificados como pertencentes a uma classe e não o foram. Deste modo a Cobertura permite reconhecer informação que foi perdida pelo classificador, esta informação é completamente invisível para a Precisão, o que leva a que normalmente sejam utilizadas em conjunto, como complementares.

$$cobertura = \frac{VP}{VP + FN}$$

5.2.4 Medida-F

A Medida-F surge da necessidade de obter um valor único representativo da qualidade do sistema analisado. Esta métrica resulta assim na combinação das duas métricas anteriormente mencionadas, a Precisão e a Cobertura, de modo a obter um valor médio.

$$F_\beta = (1 + \beta^2) \times \frac{precisão \times cobertura}{(\beta^2 \times precisão) + cobertura}$$

O parâmetro β presente na fórmula permite atribuir um peso maior a Precisão ou à Cobertura. O valor deste deve ser escolhido de acordo com a finalidade do sistema analisado, tendo em conta qual das métricas é mais importante na avaliação.

Com valores de β entre $0 \leq \beta < 1$ é atribuído um maior peso a Precisão, sendo que esse peso acentua-se quão mais próximo de 0 é o valor de β. Se for atribuído a β o valor de 0 então a importância da Cobertura será nula e o valor da Medida-F será baseado apenas na Precisão.

$$F_0 = precisão$$

Para valores de $\beta > 1$, a Cobertura terá maior peso, sendo que quando o valor atribuído a β tender para $+\infty$ o valor da Medida-F será influenciada apenas pela Cobertura.

$$\lim_{\beta \to +\infty} F_\beta = cobertura$$

Se $\beta = 1$, as medidas da Precisão e Cobertura têm o mesmo peso. A este caso particular dá-se o nome de Medida-F1[7] e o seu calculo torna-se semelhante ao da Média Harmónica[8] entre dois valores. Este é o valor de β utilizado na avaliação de resultados nesta dissertação. A fórmula da Medida-F neste cenário assume um aspeto simplificado.

$$F_1 = 2 \times \frac{precisão \times cobertura}{precisão + cobertura}$$

5.3 Resultados

Nesta secção são apresentados os resultados obtidos pelo sistema no processamento dos corpus descritos no ponto 5.1. Cada componente é analisada em separado, terminando com uma avaliação global do sistema. Os resultados são incluídos recorrendo a tabelas devidamente contextualizadas ao longo do texto.

5.3.1 Reconhecimento de Entidades Mencionadas

O REM não foi o principal foco desta dissertação, ainda assim, as alterações feitas a este módulo tiveram um impacto significativo no sistema de AS.

Foram realizados testes com as várias versões deste componente descritas no ponto 4.3.4 em dois corpus distintos. Para cada uma das versões são indicados e analisados os resultados.

SentiCorpus-PT

Os recursos *SentiCorpus-PT-gold* e *SentiCorpus-PT-clean*, já mencionados no ponto 5.1.1, foram utilizados na avaliação das várias versões deste modulo. O *SentiCorpus-PT-gold* é constituído por um total de 2416 frases, onde se encontram identificadas 3867 entidades.

[7] Do Inglês *F1-measure*.
[8] Um tipo de média apropriada quando o pretendido é uma média entre percentagens.

A avaliação foi efetuada através da comparação do *SentiCorpus-PT-gold* com o resultado do processamento do *SentiCorpus-PT-clean* pelo módulo de REM. Foram recolhidos os valores referentes às métricas mencionadas no ponto 5.2, podendo ser consultadas na tabela 5.2.

REM v1.0 A aproximação mais simples deste módulo permitiu reconhecer 1512 entidades em 1005 frases. Uma das limitações mais óbvias deste método é de facto a sua dependência do catálogo de menções a entidades. A falta de qualquer tipo de inferência relativamente a pronomes pessoais leva a que não sejam identificadas entidades num contexto composto por várias frases.

Com esta abordagem obteve-se uma Medida-F1 de 0.514. Salienta-se ainda assim a alta Precisão de 0.914, de onde se pode concluir que das entidades identificadas a esmagadora maioria destas estava correta.

REM v2.0 Ao contrário do inicialmente expectado, a utilização do sintagma nominal devolveu resultados algo desanimadores. Foram identificadas 473 entidades em 335 frases. A Medida-F foi de 0.207, embora se tenha verificado um aumento da Precisão a Cobertura desceu drasticamente com esta aproximação.

```
1                    SOCRATES 20 - PORTAS 0.
```

Listagem 28: Exemplo de frase extraída do *SentiCorpus-PT-clean*.

As razões que levaram aos resultados obtidos estão principalmente relacionadas com a natureza do corpus. A estrutura gramatical apresentada pelo texto desrespeita regularmente as regras sintáticas inerentes à língua, complicando a análise. Mesmo quando a estrutura se encontra sintaticamente correta, é extremamente simplista, não sendo composta por orações. A listagem 28 ilustra um exemplo de uma frase do corpus na qual não é possível identificar uma oração.

Devido aos fracos resultados e por se considerar que uma correção sintática automatizada do texto seria demasiado dispendiosa, esta aproximação foi deixada de lado, não chegando a ser utilizada pelo módulo de AS.

	VP	**FP**	**FN**	**Prec**	**Cob**	**F1**
REM v1.0	1382	130	2485	0.914	0.357	0.514
REM v2.0	450	23	3417	0.951	0.116	0.207
REM v3.0	1057	66	2810	0.941	0.273	0.424
REM v4.0	1095	69	2772	0.941	0.283	0.435

Tabela 5.2: Avaliação das várias versões do modulo de REM aplicado ao *SentiCorpus-PT-clean*. Valores para a Tabela de Confusão (VP, FP e FN) e valores para a Precisão, Cobertura e Medida-F1.

REM v3.0 Nesta versão foram identificadas 1123 entidades em 1005 frases. Um resultado melhor do que o obtido ao recorrer especificamente ao sintagma nominal. A maior abrangência, fornecida pelo facto de toda a oração ser considerada para a procura de entidades, permitiu duplicar o número de entidades encontradas.

Esta versão do módulo de REM obteve uma Medida-F1 de 0.424, a Precisão aumentou ligeiramente mas o grande impulso face à versão anterior vem do aumento da Cobertura, de 0.116 para 0.273. Embora os resultados desta versão tenham melhorado, ainda são inferiores aos obtidos na versão inicial do módulo. A principal justificação para os FN continua a ser a estrutura sintática simples e por vezes deficiente das frases do corpus.

REM v4.0 A última versão implementada do módulo de REM obteve uma Medida-F1 de 0.435 no SentiCorpus-PT. O pré-processamento aplicado permitiu subir para 761 o número de frases identificadas com entidade, nas quais foram assinaladas 1164 entidades. A Precisão manteve-se inalterada em relação a versão anterior. A subida ocorreu na Cobertura, passando de 0.273 para 0.283.

Esperava-se uma melhoria mais significativa nos resultados. Ainda assim foram recuperadas entidades até agora não identificadas, fundamentando a aproximação tomada.

Conclusões Os resultados finais do módulo no SentiCorpus-PT ficaram um pouco aquém do inicialmente esperado devido a alguns fatores chave. Em primeiro lugar, a limitação já assumida do catálogo de entidades. A quantidade reduzida de informação que compõe este recurso condicionou bastante a evolução do módulo. Por outro lado, a forma como as entidades fora do domínio do corpus estão anotadas influencia, de forma bastante negativa, os resultados. Casos como o ilustrado pela listagem 29 não puderam ser detetados através da aproximação seguida.

```
1     <F ID="4" ALVO="Outra Entidade" POL="0" INT="Literal">
2         <ALVOS>
3           <ALVO TIPO="NOME">Guteres</ALVO>
4         </ALVOS>
5         <VALUE>É a vida, como diria o Guteres.</VALUE>
6     </F>
```

Listagem 29: Frase do *SentiCorpus-PT-gold* com entidade anotada como "Outra Entidade".

SentiTuites-PT

A inclusão do SentiTuites-PT na avaliação foi vista como necessária de forma a permitir despistar o possível *overfitting*[9] do protótipo relativamente ao SentiCorpus-PT.

O *SentiTuites-PT-gold* é composto por um total de 5734 frases e igual número de entidades, consequência das condições de criação do recurso já mencionadas ponto 5.1.2. As restrições aplicadas ao corpus para criar o *SentiTuites-PT-clean* implicam alguma perda de significância dos valores obtidos neste corpus para outras versões, como tal apenas se verificou útil avaliar a última versão do módulo neste corpus. Tal como aconteceu com o SentiCorpus-PT, a avaliação foi efetuada através da comparação do *SentiTuites-PT-gold* com o resultado do processamento do *SentiTuites-PT-clean*. Os valores obtidos podem ser consultados na tabela 5.3.

	VP	**FP**	**FN**	**Prec**	**Cob**	**F1**
REM v4.0	3720	227	2014	0.942	0.649	0.769

Tabela 5.3: Avaliação da versão final do módulo de REM aplicado ao *SentiTuites-PT-clean*. Valores para a Tabela de Confusão (VP, FP e FN) e valores para a Precisão, Cobertura e Medida-F1.

REM v4.0 A semelhança contextual entre o SentiCorpus-PT e o SentiTuites-PT permitiu a utilização do catálogo de entidades sem alterações, sendo que os resultados foram bastante mais elevados neste último.

A ultima versão implementada do módulo identificou 3947 entidades em 3751 frases. O valor da Medida-F1 foi de 0.769, bastante superior à obtida no SentiCorpus-PT que se ficou por 0.435. A principal razão para esta diferença de valores entre os corpus deve-se às condições de criação do *SentiTuites-PT-clean*. Este não inclui tweets como a identificada pela listagem 29, presentes no *SentiCorpus-PT-clean*. O facto de apenas existir menção a uma entidade em cada frase simplifica também o processo de identificação.

Conclusões As condições de criação do *SentiTuites-PT-clean* não permitem uma comparação completamente justa entre os resultados de ambos os corpus. Servem no entanto para demonstrar que não foram criados casos específicos com vista a beneficiar os resultados obtidos no SentiCorpus-PT.

5.3.2 Análise de Sentimento

Neste ponto são analisados os resultados obtidos pelas várias versões da componente. À semelhança do que sucedeu com o módulo de REM, foram recolhidos os resultados em

[9] Adaptação excessiva das regras utilizadas a um corpus de teste, levando a que a performance do sistema nesse corpus seja inflacionada e não se reflita em novas situações.

dois corpus distintos. Estes já foram descritos no ponto 5.1.

SentiCorpus-PT

De forma a realizar uma avaliação isolada ao módulo de AS e com o objetivo de quantificar a evolução das várias versões, foi considerado um conjunto mais restrito do corpus. Este subconjunto, composto apenas pelas entradas para as quais a entidade foi corretamente identificada, é analisado pelo módulo de AS e os resultados comparados com os valores corretos correspondentes.

	Versões AS	**Frases c/ Entidade**	**Opiniões**	**Pos**	**Neg**	**Neut**
REM v1.0	1.0	931	1382	394	796	192
REM v3.0	2.0/3.0/4.0/5.0	696	1057	286	612	159
REM v4.0	6.0/7.0/8.0	724	1095	299	625	171

Tabela 5.4: Dados relativos aos subconjuntos do *SentiCorpus-PT-clean* utilizados no cálculo das métricas de avaliação para o módulo de AS.

Na tabela 5.4 constam os números relativos aos subconjuntos do corpus utilizados nesta avaliação. São indicadas as versões da componente de REM da qual resultou e as versões do módulo de AS em que foram aplicadas. Adicionalmente são indicados os valores corretos de frases com entidade e o número de opiniões total e por classe.

	Pos			**Neg**			**Neut**		
	VP	FP	FN	VP	FP	FN	VP	FP	FN
AS v1.0	75	110	319	133	57	663	151	856	41
AS v2.0	60	78	226	102	49	510	122	651	37
AS v3.0	75	102	211	154	66	458	110	556	49
AS v4.0	107	130	179	166	77	446	98	487	61
AS v5.0	109	120	177	175	77	437	98	486	61
AS v6.0	104	125	195	176	78	449	107	513	64
AS v7.0	118	159	181	201	85	424	93	445	78
AS v8.0	133	256	166	332	210	293	27	137	144

Tabela 5.5: Avaliação das várias versões do módulo de classificação de sentimento aplicado ao *SentiCorpus-PT-clean*. Valores da Tabela de Confusão por classe.

Os valores obtidos na avaliação, discriminados por classe, podem ser visualizados nas tabela 5.5 e 5.6. Uma visão mais global do desempenho do módulo é mostrado pela tabela 5.7. Nesta última são incluídos os valores correspondentes à tabela de confusão do acerto global.

AS v1.0 A primeira versão do módulo de AS classificou corretamente 359 opiniões das 1382 existentes. A simplicidade desta primeira abordagem ficou vulnerável à inexistência de adjetivos no texto analisado, algo que já era espectável.

	Pos			Neg			Neut		
	Prec	Cob	F1	Prec	Cob	F1	Prec	Cob	F1
AS v1.0	0.405	0.190	0.259	0.700	0.167	0.270	0.150	0.786	0.252
AS v2.0	0.435	0.210	0.283	0.675	0.167	0.267	0.158	0.767	0.262
AS v3.0	0.424	0.262	0.324	0.700	0.252	0.370	0.165	0.692	0.267
AS v4.0	0.451	0.374	0.409	0.683	0.271	0.388	0.168	0.616	0.263
AS v5.0	0.476	0.381	0.423	0.694	0.286	0.405	0.168	0.616	0.264
AS v6.0	0.454	0.348	0.394	0.693	0.282	0.400	0.173	0.626	0.271
AS v7.0	0.426	0.395	0.410	0.703	0.322	0.441	0.173	0.544	0.262
AS v8.0	0.342	0.445	0.387	0.613	0.531	0.569	0.165	0.158	0.161

Tabela 5.6: Avaliação das várias versões do módulo de classificação de sentimento aplicado ao *SentiCorpus-PT-clean*. Valores por classe para a Precisão, Cobertura e Medida-F1.

	VP	**FP**	**FN**	**Prec**	**Cob**	**F1**
AS v1.0	359	1023	1023	0.260	0.260	0.260
AS v2.0	284	778	773	0.267	0.269	0.268
AS v3.0	339	724	718	0.319	0.321	0.320
AS v4.0	371	694	686	0.348	0.351	0.350
AS v5.0	382	683	675	0.359	0.361	0.360
AS v6.0	387	716	708	0.351	0.353	0.352
AS v7.0	412	689	683	0.374	0.376	0.375
AS v8.0	492	603	603	0.449	0.449	0.449

Tabela 5.7: Avaliação das várias versões do módulo de classificação de sentimento aplicado ao *SentiCorpus-PT-clean*. Valores totais para a Tabela de Confusão, Precisão, Cobertura e Medida-F1.

A classe *negativo* foi a que obteve a maior Precisão, em contraste com a classe *neutro* que obteve a pior. É de salientar no entanto a discrepância entre os valores presentes no corpus para cada uma das classes. A Medida-F1 para esta versão fixou-se em 0.26.

AS v2.0 Esta versão foi a primeira a fazer uso da versão 3.0 da componente de REM. Como consequência desta alteração foi introduzida uma maior noção de contexto na associação do sentimento ao respetivo alvo. Recorda-se que no módulo de REM v3.0 foi explorado o conceito de oração. Este fator não se traduziu no entanto numa melhoria significativa dos resultados. Ocorreu um aumento no acerto relativamente às classes *positivo* e *neutro* mas uma descida no que diz respeito à classe *negativo*.

A recente limitação de contexto permitiu efetivamente melhorar a associação do sentimento ao alvo, mas o facto de apenas se considerarem adjetivos contidos nas orações deixa alguns elementos desta classe gramatical em situação inclassificável. Alguma responsabilidade atribui-se ao facto de ser um corpus composto por *tweets*, resultando numa probabilidade elevada de má construção frásica e consequente inclusão de adjetivos em locais sintaticamente incorretos.

AS v3.0 A versão 3.0 obteve uma melhoria de desempenho considerável, subindo de 0.268 para 0.320 na Medida-F1 da avaliação global. Esta subida é uma consequência direta da consideração de um novo elemento como polarizável: a classe gramatical nome. A principal razão desta subida está relacionada com o aumento geral da cobertura, mais significativo nas classes *positivo* e *negativo*. Esta situação era esperada tendo em conta que esta alteração tornou possível classificar frases que até aqui não possuíam elementos considerados polarizáveis.

AS v4.0 Um pouco à semelhança do que aconteceu na versão anterior, a versão 4.0 desta componente considerou a polaridade de uma terceira classe gramatical: o verbo. Esta alteração elevou, mais uma vez, a Cobertura de uma forma geral, embora com o registo de uma pequena descida para a classe *neutro*. Paralelamente ocorreu um aumento significativo da Precisão para a classe *positivo*. Estes resultados elevaram a Medida-F1 global para 0.350, demonstrando efetivamente a existência de informação válida sobre sentimento associada a esta classe gramatical.

AS v5.0 A versão 5.0 da componente conseguiu uma pequena melhoria na Medida-F1 global, subindo para 0.360. Embora não muito expressiva, a alteração conseguiu ter um impacto positivo em todas as classes. Este resultado demonstra o efeito corretivo da alteração, de certa forma já esperado, confirmando assim o peso da negação na língua. O impacto dos restantes modificadores de intensidade do sentimento foi positivo, um pouco menos expressivo que o causado pela inclusão da negação mas ainda assim claramente benéfico.

AS v6.0 Nesta versão não foram realizadas alterações diretas ao módulo de AS. A diferença nos resultados é causada principalmente pelo aumento do número de frases consideradas na avaliação, consequência das alterações ao módulo de REM. Este crescimento da amostra considerada na avaliação levou à inclusão de novos casos, muitos dos quais não foram corretamente classificados, refletindo-se num impacto negativo nas métricas de acerto. Por outro lado, o pré-processamento realizado ao texto inicial levou a que a classe gramatical identificada pelo módulo de AM se alterasse em alguns casos, modificando a avaliação subsequente efetuada pela componente de AS.

AS v7.0 A versão 7.0 do módulo de AS procurou atribuir uma importância diferente a cada classe gramatical considerada polarizável. Esta alteração permitiu classificar corretamente casos em que, pela inclusão de uma maior quantidade de nomes ou verbos na frase, a polaridade do sentimento era incorretamente detetada.

Como já referido aquando da descrição desta versão do módulo, a relação de importância entre as três classes gramaticais que evidenciou melhores resultados, do mais valioso para o menos, foi:

1. Adjetivos - 1.0
2. Nomes - 0.6
3. Verbos - 0.2

Os números presentes na lista correspondem aos pesos aplicados aos elementos da correspondente classe gramatical. Como já mencionado, seria necessária uma investigação um pouco mais extensa no sentido de obter os valores ótimos para cada componente morfológico. Ainda assim, foram realizados testes com várias combinações em que as prioridades foram parcial e completamente invertidas, sendo que a combinação que obteve melhores resultados foi a previamente listada.

É de salientar que os resultados obtidos possuem uma elevada dependência do corpus em análise. Seria necessário investigar efetivamente qual o impacto desta aproximação numa maior coleção de textos, sendo espectável a necessidade de alterações de acordo com o domínio e origem do conteúdo a classificar. Fica no entanto demonstrado o valor da aproximação, revelando-se algo a ter em conta nas futuras afinações do sistema.

Esta alteração apenas não melhorou o acerto para a classe *neutro*, revelando-se positiva para as restantes. A Medida-F1 global do módulo subiu de 0.352 para 0.375, consequência de uma subida paralela da Precisão e da Cobertura.

AS v8.0 A última versão implementada do módulo de AS trouxe os benefícios esperados, mas também alguns problemas. Como já referido anteriormente no capítulo 4, o objetivo da utilização do dicionário de sinónimos, descrito no ponto 4.4.2, foi aumentar a capacidade de classificação do sistema. Procurou-se assim contornar as limitações do léxico de sentimentos explorando a relação de sinonímia entre as palavras.

A primeira consequência desta aproximação é precisamente o esperado: um aumento considerável da Cobertura global da componente.

	Total Elementos	**Classificados v7.0**	**Classificados v8.0**
Adjetivos	1196	610	788
Verbos	3726	466	1751
Nomes	3402	296	1067

Tabela 5.8: Aumento no número de elementos que o módulo de AS conseguiu classificar com a utilização do Dicionário de Sinónimos.

A tabela 5.8 mostra o aumento da capacidade de classificação de cada elemento polarizável. Este crescimento é especialmente evidente nas classes gramaticais dos nomes e dos verbos, aumentando em mais de quatro vezes a capacidade de classificação nestes casos.

Pelo lado negativo, a utilização deste método implica inevitavelmente um aumento de erros na atribuição de polaridade. Estes erros são normalmente gerados por relações de

sinonímia incorretamente definidas ou válidas para contextos diferentes daquele em que se insere o elemento original. Como consequência é observada uma diminuição geral da Precisão em todas as classes. Ainda assim, o efeito positivo nos resultados é notório, apoiado principalmente pelo aumento da Cobertura como já mencionado. Destaca-se a subida considerável da Medida-F1 global de 0.375 para 0.449.

Conclusões É indiscutível a evolução positiva, demonstrada pelas consecutivas melhorias na avaliação, mas continuam no entanto evidentes algumas limitações ainda por contornar. O exemplo contido na listagem 28, mencionado a propósito da análise do módulo de REM, mantém-se inclassificável através das aproximações tomadas. O domínio do corpus analisado é bastante propício à existência de ironia e expressões idiomáticas, fraqueza identificada no sistema e excelente candidata a futuros desenvolvimentos. Estas são algumas das questões que impedem um melhor desempenho do sistema. Admite-se que ao encontrar solução para estes problemas, em conjunto com a minimização das limitações demonstradas pelos recursos utilizados, o sistema atinja resultados bastante mais elevados.

SentiTuites-PT

À semelhança do que aconteceu com a avaliação do módulo de REM, a inclusão do SentiTuites-PT na avaliação foi vista como necessária de forma a permitir despistar o possível *overfitting* do protótipo relativamente ao SentiCorpus-PT.

	Versões AS	**Frases c/ Entidade**	**Opiniões**	**Pos**	**Neg**	**Neut**
REM v4.0	8.0	3720	3720	358	2348	1014

Tabela 5.9: Dados relativos aos subconjunto do *SentiTuites-PT-clean* utilizado no cálculo das métricas de avaliação para o módulo de AS v8.0.

Tal como sucedeu com a avaliação no SentiCorpus-PT, apenas foi considerado um subconjunto de elementos do *SentiTuites-PT-clean*. Este subconjunto corresponde ao número de entradas onde o alvo de sentimento foi corretamente identificado pelo módulo de REM. A tabela 5.9 contém os números relativos a este subconjunto.

	Pos			**Neg**			**Neut**		
	VP	FP	FN	VP	FP	FN	VP	FP	FN
AS v8.0	151	907	207	1226	619	1122	301	568	713

Tabela 5.10: Avaliação da v8.0 do módulo de classificação de sentimento aplicado ao *SentiTuites-PT-clean*. Valores da Tabela de Confusão por classe.

Dadas as limitações inerentes à criação do corpus, mencionadas no ponto 5.1.2, optou-se por incluir apenas os resultados da última versão do módulo de AS. O resultado da classificação realizada no subconjunto do *SentiTuites-PT-clean* foi comparada face ao mesmo subconjunto de elementos do *SentiTuites-PT-gold*. Os resultados podem ser visualizados nas tabelas 5.10, 5.11 e 5.12.

	Pos			Neg			Neut		
	Prec	Cob	F1	Prec	Cob	F1	Prec	Cob	F1
AS v8.0	0.143	0.422	0.213	0.664	0.522	0.585	0.346	0.297	0.320

Tabela 5.11: Avaliação da v8.0 do módulo de classificação de sentimento aplicado ao *SentiTuites-PT-clean*. Valores por classe para a Precisão, Cobertura e Medida-F1.

	VP	FP	FN	Prec	Cob	F1
AS v8.0	1678	2094	2042	0.445	0.451	0.448

Tabela 5.12: Avaliação da v8.0 do módulo de classificação de sentimento aplicado ao *SentiTuites-PT-clean*. Valores totais para a Tabela de Confusão, Precisão, Cobertura e Medida-F1.

AS v8.0 Demonstrando exatamente o pretendido, a independência do corpus, os resultados obtidos pelo módulo no SentiTuites-PT foram bastante semelhantes. Embora o SentiTuites-PT seja um corpus constituído por *tweets*, levantando naturalmente os problemas inerentes a este tipo de fonte, as aproximações realizadas pelo sistema mantiveram-se válidas, demonstrando assim o seu valor.

A v8.0 do módulo de AS obteve uma Medida-F1 global de 0.448 no SentiTuites-PT, face aos 0.449 obtidos no SentiCorpus-PT.

Conclusões Através da utilização do SentiTuites-PT nas mesmas condições em que foi utilizado o SentiCorpus-PT demonstra-se a independência das aproximações escolhidas para o módulo de AS.

Aprendizagem Automática

Torna-se pertinente incluir uma comparação entre os resultados do protótipo desenvolvido e os obtidos através de uma aproximação completamente diferente. A aproximação referida já foi apresentada no ponto 3.2 e consiste em recorrer a AA para realizar a classificação de sentimento.

Embora não tenham sido realizados testes muito complexos, admitem-se como suficientes para justificar as razões pelas quais se optou por um sistema baseado em regras em oposição à AA.

O algoritmo escolhido para os testes foi o *Naive Bayes Multinomial*[10] implementado recorrendo a ferramenta *Weka*, já descrita no ponto 2.2.3. Foram utilizados ambos os corpus, SentiTuites-PT e SentiCorpus-PT, sem nenhum tratamento adicional específico para este efeito. Salienta-se que apenas foi avaliada a classificação de sentimento, assim para eliminar a influência do REM recorreu-se aos mesmos subconjuntos dos corpus já utilizados na

[10] Algoritmo estatístico simples baseado na aplicação do teorema de *Bayes*. Este algoritmo obtém na maioria dos casos resultados satisfatórios tendo em conta a sua simplicidade.

avaliação da última versão do módulo de AS.

	Pos			Neg			Neut		
N-Gram	Prec	Cob	F1	Prec	Cob	F1	Prec	Cob	F1
1	0.676	0.391	0.495	0.702	0.900	0.789	0.118	0.036	0.055
1 e 2	0.617	0.453	0.523	0.711	0.858	0.777	0.091	0.036	0.051
1, 2 e 3	0.629	0.438	0.516	0.716	0.858	0.781	0.300	0.161	0.209

Tabela 5.13: Avaliação da classificação de sentimento no SentiCorpus-PT com *Naive Bayes Multinomial*. Valores por classe para a Precisão, Cobertura e Medida-F1 recorrendo a *unigrams*, *bigrams* e *trigrams*.

N-Gram	**Prec**	**Cob**	**F1**
1	0.632	0.679	0.636
1 e 2	0.620	0.667	0.635
1, 2 e 3	0.649	0.677	0.653

Tabela 5.14: Avaliação da classificação de sentimento no SentiCorpus-PT com *Naive Bayes Multinomial*. Valores totais para a Precisão, Cobertura e Medida-F1 recorrendo a *unigrams*, *bigrams* e *trigrams*.

As tabelas 5.13 e 5.14 contêm os resultados da classificação do SentiCorpus-PT. Foram considerados *unigrams*, *bigrams* e *trigrams* na avaliação. A divisão do texto em *tokens* e cálculo de frequências foi realizado recorrendo ao filtro *StringToWordVector* já implementado no *Weka*. Para definir o particionamento do corpus para treino e avaliação foi utilizado o método da Validação Cruzada[11] 10-*fold*, tido como comum na avaliação de sistemas de classificação estatísticos com base num conjunto de dados.

	Pos			Neg			Neut		
Treino/Class	Prec	Cob	F1	Prec	Cob	F1	Prec	Cob	F1
ST/SC	0.259	0.438	0.326	0.667	0.430	0.523	0.129	0.196	0.156
SC/ST	0.253	0.083	0.125	0.623	0.975	0.760	0.400	0.001	0.002

Tabela 5.15: Avaliação da classificação de sentimento com treino no SentiCorpus-PT e classificação no SentiTuites-PT e com treino no SentiTuites-PT; e avaliação no SentiCorpus-PT. Valores por classe para a Precisão, Cobertura e Medida-F1 recorrendo a *unigrams* com *Naive Bayes Multinomial*.

As tabelas 5.15 e 5.16 listam os resultados de uma avaliação com características ligeiramente diferentes. Neste caso todo o treino foi realizado num dos corpus e a classificação levada a cabo no outro. Ambos os corpus ocuparam posições de treino e de classificado. Desta vez apenas se utilizaram *unigrams* sendo as restantes condições semelhantes à avaliação anterior.

[11] Técnica utilizada em que o conjunto de dados é particionado e avaliado iterativamente trocando os conjuntos de treino e classificação. Uma média é posteriormente calculada de forma a normalizar os resultados das várias rondas.

Treino/Class	Prec	Cob	F1
ST/SC	0.507	0.407	0.434
SC/ST	0.524	0.611	0.482

Tabela 5.16: Avaliação da classificação de sentimento com treino no SentiCorpus-PT e classificação no SentiTuites-PT e com treino no SentiTuites-PT; e avaliação no SentiCorpus-PT. Valores totais para a Precisão, Cobertura e Medida-F1 recorrendo a *unigrams* com *Naive Bayes Multinomial*.

Conclusões Ao analisar os resultados obtidos através da AA, se nos debruçarmos sobre a primeira abordagem, é necessário reconhecer a superioridade. A Medida-F1 de 0.653 obtida na classificação do SentiCorpus-PT é claramente superior aos 0.449 obtidos pelo protótipo. Porém, é pertinente relembrar uma das fragilidades da AA: a necessidade de conteúdo manualmente anotado para treino. Ao limitar a informação disponível para treino do modelo, a sua incapacidade perante novas situações fica imediatamente exposta. Para de certa forma simular esta fragilidade foi levado a cabo o segundo conjunto de testes. Ao treinar num corpus ligeiramente diferente, ainda que pertencente ao mesmo domínio, essa fragilidade revelou-se obtendo resultados bastante mais baixos.

Um sistema baseado em regras como aquele aqui apresentado possui sem dúvida limitações, algumas das quais já mencionadas ao longo desta avaliação, no entanto é totalmente independente da existência de conteúdo de treino. Acredita-se que com mais investigação e afinação os resultados obtidos serão amplamente superiores aos demonstrados pelos sistemas de AA.

5.3.3 Sistema

Neste ponto são avaliados os resultados obtidos pelo sistema no corpora em estudo. Tendo em conta o detalhe presente na avaliação dos vários módulos, e o facto do processamento do sistema corresponder diretamente ao trabalho desempenhado pelas suas componentes, será apenas realizada uma apreciação global dos resultados.

	Pos			**Neg**			**Neut**		
	VP	FP	FN	VP	FP	FN	VP	FP	FN
v1.0	75	123	715	133	66	2398	151	964	399
v2.0	60	82	730	102	58	2429	122	706	428
v3.0	75	108	715	154	83	2377	110	602	440
v4.0	107	144	683	166	98	2365	98	519	452
v5.0	109	134	681	175	99	2356	98	517	452
v6.0	104	139	686	176	99	2355	107	548	443
v7.0	118	178	672	201	108	2330	93	472	457
v8.0	133	287	657	332	246	2199	27	139	523

Tabela 5.17: Avaliação das várias versões do sistema aplicado ao *SentiCorpus-PT-clean*. Valores da Tabela de Confusão por classe.

	Pos			Neg			Neut		
	Prec	Cob	F1	Prec	Cob	F1	Prec	Cob	F1
AS v1.0	0.379	0.095	0.152	0.668	0.053	0.097	0.135	0.275	0.181
AS v2.0	0.423	0.076	0.129	0.638	0.040	0.076	0.147	0.222	0.177
AS v3.0	0.410	0.095	0.154	0.650	0.061	0.111	0.154	0.200	0.174
AS v4.0	0.426	0.135	0.206	0.629	0.066	0.119	0.159	0.178	0.168
AS v5.0	0.449	0.138	0.211	0.639	0.069	0.125	0.159	0.178	0.168
AS v6.0	0.428	0.132	0.201	0.640	0.070	0.125	0.163	0.195	0.178
AS v7.0	0.399	0.149	0.217	0.650	0.079	0.142	0.165	0.169	0.167
AS v8.0	0.317	0.168	0.220	0.574	0.131	0.214	0.163	0.049	0.075

Tabela 5.18: Avaliação das várias versões do sistema aplicado ao *SentiCorpus-PT-clean*. Valores por classe para a Precisão, Cobertura e Medida-F1.

	VP	**FP**	**FN**	**Prec**	**Cob**	**F1**
v1.0	359	1,153	3,512	0.237	0.093	0.133
v2.0	284	846	3,587	0.251	0.073	0.114
v3.0	339	793	3,532	0.299	0.088	0.136
v4.0	371	761	3,500	0.328	0.096	0.148
v5.0	382	750	3,489	0.337	0.099	0.153
v6.0	387	786	3,484	0.330	0.100	0.153
v7.0	412	758	3,459	0.352	0.106	0.163
v8.0	492	672	3,379	0.423	0.127	0.195

Tabela 5.19: Avaliação das várias versões do sistema aplicado ao *SentiCorpus-PT-clean*. Valores totais para a Tabela de Confusão, Precisão, Cobertura e Medida-F1.

As tabelas 5.17, 5.18 e 5.19 contêm a avaliação dos resultados obtidos pelo sistema quando aplicado ao SentiCorpus-PT. Como é natural, a evolução destes é coerente com a avaliação do módulo de AS. Os valores mais baixos na Medida-F1 global, quando comparados com os obtidos pelo módulo de AS, devem-se à influência da componente de REM. Se o alvo do sentimento estiver incorreto o resultado é considerado imediatamente como errado, facto que se reflete negativamente na avaliação.

	Pos			Neg			Neut		
	VP	FP	FN	VP	FP	FN	VP	FP	FN
v8.0	151	964	401	1,226	723	2317	301	636	1338

Tabela 5.20: Avaliação da versão 8.0 do sistema aplicado ao *SentiTuites-PT-clean*. Valores da Tabela de Confusão por classe.

	Pos			Neg			Neut		
	Prec	Cob	F1	Prec	Cob	F1	Prec	Cob	F1
AS v8.0	0.135	0.274	0.181	0.629	0.346	0.446	0.321	0.184	0.234

Tabela 5.21: Avaliação da versão 8.0 do sistema aplicado ao *SentiTuites-PT-clean*. Valores por classe para a Precisão, Cobertura e Medida-F1.

As tabelas 5.20, 5.21 e 5.22 apresentam a avaliação do sistema calculada sobre o processa-

	VP	FP	FN	Prec	Cob	F1
v8.0	1,678	2,323	4,056	0.419	0.293	0.345

Tabela 5.22: Avaliação da versão 8.0 do sistema aplicado ao *SentiTuites-PT-clean*. Valores totais para a Tabela de Confusão, Precisão, Cobertura e Medida-F1.

mento do SentiTuites-PT. A diferença de valores face ao processamento realizado sobre o SentiCorpus-PT justifica-se com as condições de adaptação do corpus. O facto de apenas terem sido consideradas frases com uma entidade leva a que o desempenho do módulo de REM esteja inflacionado, elevando a Medida-F1 global do sistema.

5.4 Conclusões

A avaliação realizada sobre o sistema procurou justificar as opções tomadas durante a sua construção e desenvolvimento. Ficou demonstrado através dos resultados obtidos o valor acrescentado de cada aproximação. No entanto, a mesma avaliação revelou fraquezas em todo o modelo, merecedoras de atenção e estudo futuro, no sentido de serem atingidos resultados capazes de rivalizar com os sistemas mais atuais no mercado.

Por entre os principais problemas detetados estão os erros de escrita e construção frásica, bastante comuns nas fontes de onde provêm os corpus processados. A deteção de ironia e reconhecimento de expressões idiomáticas é outro ponto que contribuiu para a taxa de erro do sistema. A fragilidade dos recursos utilizados, nomeadamente o léxico de sentimentos e o dicionário de sinónimos, tiveram também a sua contribuição negativa na avaliação. Finalmente, a simplicidade da aproximação utilizada na componente de REM influencia largamente os resultados globais do sistema. Esta última componente merece efetivamente ser alvo de um extenso trabalho futuro.

Contudo, crê-se ter ficado demonstrada a capacidade da aproximação escolhida e justificada a opção por um sistema baseado em regras.

Capítulo 6

Conclusões

O crescimento acentuado das redes sociais *online* nos últimos anos é o resultado da aceitação global deste tipo de serviços. Atualmente, "*twittar*", "*postar no face*" ou "*fazer like*" são expressões que integram o nosso quotidiano e que facilmente encontramos em campanhas publicitárias bem como média em geral. A sociedade contemporânea apoia-se cada vez mais nestes serviços como meio de comunicação, deixando por vezes a sua utilização de ser classificada como entretenimento, passando efetivamente a ser uma ferramenta de trabalho. A comodidade, facilidade e ausência de custos associadas a estes torna-os de certa forma incontornáveis, sendo por isso difícil justificar a sua não utilização.

O alarido criado pelas redes sociais não tem passado despercebido ao mundo empresarial. Numa realidade económica atualmente difícil, saber o que os clientes pensam, os seus gostos e interesses, pode significar estar um passo a frente da concorrência. Toda esta informação existe e está disponível, porém a sua quantidade e formato elevam a complexidade do seu aproveitamento. Uma aproximação manual está cada vez mais fora de questão, sendo necessária a criação de ferramentas capazes de extrair significado de todo este conteúdo. Desde a década de 80 que a comunidade científica se debruça sobre o problema, no entanto é a partir de 2001 que a temática recebe maior atenção, respondendo assim à pressão do mercado que anseia por resultados.

6.1 Balanço Final

Nesta dissertação procura-se, numa fase inicial, ilustrar o estado atual dos sistemas de AS, descrevendo as abordagens mais comuns ao problema. É realizada uma contextualização ao tema com o objetivo de facilitar a compreensão e dar uma visão do trabalho

realizado até à atualidade. Apresenta-se a arquitetura de um sistema capaz de classificar automaticamente opiniões sobre entidades em texto, sendo posteriormente analisados os seus resultados e comparados com outras aproximações.

O sistema proposto é constituído por três módulos: Analisador Morfossintático, Reconhecedor de Entidades Mencionadas e o Analisador de Sentimento.

O Analisador Morfossintático realiza um pré-processamento do texto a classificar, enriquecendo este com informação relativamente à morfologia e sintaxe dos vários elementos que o compõem. O Reconhecedor de Entidades Mencionadas procura identificar menções a entidades no texto, sendo estas os alvos sobre os quais o sentimento poderá incidir. Finalmente o Analisador de Sentimento realiza uma classificação de polaridade da opinião sobre a entidade encontrada, seguindo uma aproximação de OS recorrendo à utilização de um léxico de sentimentos e de um dicionário de sinónimos.

Como já foi dito anteriormente, o modelo apresentado segue uma metodologia de OS, tornando-se relevante situar a solução apresentada no panorama atual apresentado no capítulo 3.

Os sistemas baseados numa metodologia de OS, como os apresentados no ponto 3.3, recorrem normalmente a um grupo de regras pré-determinado para realizar a classificação. É também bastante comum a utilização de léxicos ou outro tipo de catálogos durante o processo. A solução proposta enquadra-se neste cenário, fazendo uso de ambas as técnicas.

Um pouco mais em particular, é possível identificar alguns pontos em comum com o sistema SentiCorr, apresentado no ponto 3.3.1. Tal como no SentiCorr, a solução proposta recorre à extração de informação morfológica como forma de enriquecer o texto numa fase de pré-processamento. A interpretação de modificadores de intensidade adotada no modelo proposto pode também ser encontrada no SentiCorr, sendo no entanto encarada de uma forma mais simples neste sistema. O projeto PIRPO, apresentado no ponto 3.3.2, inclui também algumas técnicas em comum com o modelo desta dissertação. Salienta-se a utilização de um léxico de sentimentos como recurso auxiliar, de forma semelhante ao sistema apresentado, e uma ontologia de conceitos. O contexto de utilização desta última pode considerar-se de certa forma semelhante aquele em que foi inserido o catálogo de entidades, recurso utilizado no módulo de REM da solução proposta, sendo o seu propósito ajudar na identificação dos alvos do sentimento.

Existe uma característica diferenciadora na abordagem desta dissertação relativamente aos sistemas de OS apresentados, a total orientação para o alvo sobre o qual incide o sentimento. Um pouco à semelhança do sistema baseado num modelo de AA, referido no ponto 3.2.2, a solução proposta associa o sentimento a um alvo, considerando irrelevantes opiniões que não sejam direcionadas para algo ou alguém.

Esta dissertação propôs-se a apresentar uma solução modular capaz de identificar e classificar menções a entidades em texto. Esse objetivo crê-se ter sido atingido através do sistema apresentado. O seu carácter modular permite uma evolução paralela e independente das

suas várias componentes, enquanto a sua natureza de OS lhe confere uma maior adaptabilidade a novos cenários, sem a necessidade do treino adicional inerente aos sistemas de AA.

Na avaliação o sistema obteve uma Medida-F1 final de 0.195, e uma Precisão de 0.423 no corpus utilizado. Este valor refere-se ao acerto no alvo e no respetivo sentimento que incide sobre este. A análise dos resultados demonstrou várias fragilidades no modelo, ainda assim a comparação realizada com uma aproximação de AA simples, demonstra a existência de algumas potencialidades no modelo adotado. Como é comum num sistema de OS, a qualidade das regras utilizadas dita a performance da sua classificação. Este é simultaneamente o ponto forte e fraco desta metodologia. É um ponto forte na medida em que um bom conjunto de regras confere uma classificação exata sem ser necessária uma fase de treino, como sucede na AA. Analogamente, é um ponto fraco tendo em conta a dificuldade em definir esse conjunto de regras, podendo ser facilmente superado por um sistema de AA treinado para o domínio em que é aplicado. O modelo apresentado carece de mais regras por forma a tornar a sua vertente de OS num ponto forte, sendo que na implementação apresentada pode facilmente ser ultrapassado.

A ideia que deu origem a esta dissertação era um pouco mais ambiciosa do que o que acabou por ser o produto final. Uma das vertentes que ficou bastante aquém do espectado foi a deteção de entidades, um pouco negligenciada como consequência da sua complexidade e de um compromisso temporal previamente assumido, carecia efetivamente de mais investigação. A classificação de sentimento sempre foi tida como a componente prioritária do sistema, no entanto também nesta vertente o produto final divergiu um pouco do previsto. A comparação entre entidades, a deteção de ironia e a resolução de anáforas faziam parte dos objetivos iniciais que eventualmente tiveram de ser deixados de parte, sendo agora incluídos no trabalho futuro a realizar sobre o tema.

6.2 Trabalho Futuro

Tendo em conta o conjunto de fragilidades detetadas durante a avaliação do sistema proposto, e a complexidade inerente ao tema, existem várias vertentes merecedoras de trabalho e investigação.

O funcionamento da componente que realiza o enriquecimento morfossintático é bastante penalizado quando esta se depara com texto incorretamente formado, vulgarmente presente nas fontes em análise. A inclusão de um pré-processamento corretivo, ou a eventual utilização de um classificador morfossintático específico para este tipo de texto são casos em aberto passiveis de atenção.

O reconhecimento de entidades mencionadas não mereceu a atenção devida ao longo desta dissertação, devendo ser alvo de algum trabalho por forma a melhorar o seu desempenho. O alargamento do catálogo de entidades utilizado será um bom investimento, bem como a inclusão de mais informação relevante para cada tipo de menção. O mecanismo de deteção

de menções poderá também ser melhorado, apoiando-se possivelmente um pouco mais na sintaxe e nas regras de construção frásica.

Quanto ao analisador de sentimentos, o principal trabalho a realizar estará no aumento da fiabilidade dos recursos utilizados. Uma maior completude do léxico de sentimentos, tanto ao nível do número de palavras classificadas como também ao nível da informação sobre a polaridade, por forma a permitir uma classificação mais precisa. A exploração das relações de sinonímia é também um ponto passível de melhoria, aumentando não só a sua capacidade como explorando outras relações tais como a antonímia. A deteção de subjetividade é outro ponto onde é necessário um estudo mais aprofundado, permitindo melhorar a classificação do sistema na classe neutro, onde este possui os piores resultados. Finalmente, a inclusão de novas regras de classificação que permitam a deteção de ironia através de expressões idiomáticas e identificação de sentimento através da comparação de entidades.

A adoção de uma aproximação de AA na associação do alvo ao sentimento ou talvez na deteção de ironia, dando assim um carácter híbrido ao sistema, é efetivamente um possível caminho a explorar. A inclusão desta metodologia poderia permitir colmatar algumas das fragilidades atualmente presentes no sistema, complementando de alguma forma as regras já empregues.

A sociedade da informação reconhece finalmente o potencial da AS. Saber como as pessoas se sentem hoje permitirá ir de encontro às suas necessidades de amanhã.

Bibliografia

[1] *Dicionário editora da língua portuguesa 2013*, 2012, p. 1744.

[2] R. Acharyya, *A new approach for blind source separation of convolutive sources: Wavelet based separation using shrinkage function*, VDM, Verlag Dr. Müller, 2008.

[3] J. Allen, *Natural language understanding*, (1987).

[4] M. Aronoff and K. Fudeman, *What is morphology*, vol. 6, Wiley-Blackwell, 2010.

[5] L. Barbosa and J. Feng, *Robust sentiment detection on twitter from biased and noisy data*, Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, 2010, pp. 36–44.

[6] E. Bick, *The parsing system"palavras": Automatic grammatical analysis of portuguese in a constraint grammar framework*, vol. 202, Aarhus University Press Aarhus,, Denmark, 2000.

[7] A. Branco and J. Silva, *Evaluating solutions for the rapid development of state-of-the-art pos taggers for portuguese*, Proceedings of the 4th Language Resources and Evaluation Conference (LREC), 2004, pp. 507–510.

[8] E. Brill and R.J. Mooney, *An overview of empirical natural language processing*, AI magazine **18** (1997), no. 4, 13.

[9] J.G. Carbonell, *Subjective understanding: computer models of belief systems.*, Tech. report, DTIC Document, 1979.

[10] P. Carvalho, L. Sarmento, J. Teixeira, and M.J. Silva, *Liars and saviors in a sentiment annotated corpus of comments to political debates*, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers, vol. 2, 2011, pp. 564–568.

[11] M. Chaves, L. de Freitas, M. Souza, and R. Vieira, *Pirpo: An algorithm to deal with polarity in portuguese online reviews from the accommodation sector*, Natural Language Processing and Information Systems (2012), 296–301.

[12] M. Chaves and C. Trojahn, *Towards a multilingual ontology for ontology-driven content mining in social web sites*, Proceedings of the ISWC 2010 Workshops, Volume I, 1st International Workshop on Cross-Cultural and Cross-Lingual Aspects of the Semantic Web, 2010.

[13] P. Chesley, B. Vincent, L. Xu, and R.K. Srihari, *Using verbs and adjectives to automatically classify blog sentiment*, Training **580** (2006), no. 263, 233.

[14] comScore/the Kelsey group, *Online consumer-generated reviews have significant impact on offline purchase behavior*, Press Release, November 2007, http://www.comscore.com/press/release.asp?press=1928.

[15] F. Cozman and I. Cohen, *Risks of semi-supervised learning: How unlabeled data can degrade performance of generative classifiers*, Semi-Supervised Learning **4** (2006), 57–72.

[16] N.B. Ellison et al., *Social network sites: Definition, history, and scholarship*, Journal of Computer-Mediated Communication **13** (2007), no. 1, 210–230.

[17] A. Esuli and F. Sebastiani, *Sentiwordnet: A publicly available lexical resource for opinion mining*, Proceedings of LREC, vol. 6, Citeseer, 2006, pp. 417–422.

[18] Y. Freund and R. Schapire, *A desicion-theoretic generalization of on-line learning and an application to boosting*, Computational learning theory, Springer, 1995, pp. 23–37.

[19] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger, *Pulse: Mining customer opinions from free text*, Advances in Intelligent Data Analysis VI (2005), 741–741.

[20] A. Go, R. Bhayani, and L. Huang, *Twitter sentiment classification using distant supervision*, CS224N Project Report, Stanford (2009), 1–12.

[21] B.J. Grosz and F.N.C. Pereira, *Natural language processing*, MIT Press, 1994.

[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, *The weka data mining software: an update*, ACM SIGKDD Explorations Newsletter **11** (2009), no. 1, 10–18.

[23] Y.B. Hillel, *Language and information*, 1964.

[24] J.A. Horrigan, *Online shopping*, Pew Internet & American Life Project Report **36** (2008).

[25] A. Java, X. Song, T. Finin, and B. Tseng, *Why we twitter: understanding microblogging usage and communities*, Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, ACM, 2007, pp. 56–65.

[26] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, *Target-dependent twitter sentiment classification*, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, 2011, pp. 151–160.

[27] K.S. Jones, *Natural languages processing: a historical review*, University of Cambridge (2001).

[28] I. Maks and P. Vossen, *A verb lexicon model for deep sentiment analysis and opinion mining applications*, Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011), 2011, pp. 10–18.

[29] G.B. Matthews, *Accidental unities*, Language and Logos (1982), 223–240.

[30] D. Maynard, H. Saggion, M. Yankova, K. Bontcheva, and W. Peters, *Natural language technology for information integration in business intelligence*, Business Information Systems, Springer, 2007, pp. 366–380.

[31] G.A. Miller et al., *Wordnet: a lexical database for english*, Communications of the ACM **38** (1995), no. 11, 39–41.

[32] S. Moreira, D. Batista, P. Carvalho, F.M. Couto, and M.J. Silva, *Power-politics ontology for web entity retrieval*.

[33] Mário Mourão and José Saias, *Bclaas: implementação de uma base de conhecimento linguístico as-a-service*, Actas das 3ªs Jornadas de Informática da Universidade de Évora - JIUE'2013, Escola de Ciências e Tecnologia da Universidade de Évora, 2013.

[34] D. Nadeau and S. Sekine, *A survey of named entity recognition and classification*, Lingvisticae Investigationes **30** (2007), no. 1, 3–26.

[35] B. Pang and L. Lee, *Opinion mining and sentiment analysis*, Now Pub, 2008.

[36] L. Rainie and J. Horrigan, *Election 2006 online*, Pew Internet & American Life Project Report (2007).

[37] A. Ratnaparkhi et al., *A maximum entropy model for part-of-speech tagging*, Proceedings of the conference on empirical methods in natural language processing, vol. 1, Philadelphia, PA, 1996, pp. 133–142.

[38] José Saias and Paulo Quaresma, *Semantic networks and spreading activation process for qa improvement on text answers*, Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology - STIL2011 (Cuiabá, Mato Grosso, Brasil), 2011, ISSN: 2175-6201.

[39] A. Santos, H. Oliveira, C. Ramos, and N. Marques, *A bootstrapping algorithm for learning the polarity of words*, Computational Processing of the Portuguese Language (2012), 229–234.

[40] H. Schmid, *Probabilistic part-of-speech tagging using decision trees*, Proceedings of international conference on new methods in language processing, vol. 12, Manchester, UK, 1994, pp. 44–49.

[41] F. Sebastiani, *Machine learning in automated text categorization*, ACM computing surveys (CSUR) **34** (2002), no. 1, 1–47.

[42] T.J. Sejnowski, *Unsupervised learning: foundations of neural computation*, The MIT press, 1999.

[43] Mário J. Silva, Paula Carvalho, and Luís Sarmento, *Building a sentiment lexicon for social judgement mining*, Lecture Notes in Computer Science, vol. 7243, Springer Berlin Heidelberg, 2012.

[44] M. Souza, R. Vieira, D. Busetti, R. Chishman, and I.M. Alves, *Construction of a portuguese opinion lexicon from multiple resources*, STIL (2011).

[45] E. Tognini-Bonelli, *Corpus linguistics at work*, vol. 6, John Benjamins Publishing Co, 2001.

[46] E. Tromp, *Multilingual sentiment analysis on social media*, Master's Theisis. Department of Mathematics and Computer Science, Eindhoven University of Technology (2011).

[47] E. Tromp and M. Pechenizkiy, *Graphbased n-gram language identification on short texts*, Proc. 20th Machine Learning conference of Belgium and The Netherlands, 2011, pp. 27–34.

[48] Erik Tromp and Mykola Pechenizkiy, *Senticorr: Multilingual sentiment analysis of personal correspondence*, Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, IEEE, 2011, pp. 1247–1250.

[49] R.D. Van Valin, *An introduction to syntax*, Cambridge Univ Pr, 2001.

[50] D.L. Waltz, *An english language question answering system for a large relational database*, Communications of the ACM **21** (1978), no. 7, 526–539.

[51] M. Wiegand, A. Balahur, B. Roth, D. Klakow, and A. Montoyo, *A survey on the role of negation in sentiment analysis*, Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, Association for Computational Linguistics, 2010, pp. 60–68.

[52] T. Winograd, *Understanding natural language*, Cognitive Psychology **3** (1972), no. 1, 1–191.

[53] W.A. Woods, *Lunar rocks in natural english: Explorations in natural language question answering*, Linguistic structures processing **5** (1977), 521–569.