UNIVERSIDADE DA BEIRA INTERIOR
Engenharia

# Análise de sentimento em texto no domínio económico
# (Versão Final Após Defesa)

Mvita Zankulu

Dissertação para obtenção do Grau de Mestre em
Engenharia informática
(2º ciclo de estudos)

Orientador: Prof. Doutor João Paulo Cordeiro

Covilhã, Dezembro de 2018

# Dedicatória

Dedico este trabalho a Deus e a minha família pelo total apoio nesta formação.

# Agradecimento

Os meus agradecimentos dirigem-se:

a Deus por ter-me dado saúde, vontade e força para enfrentar e superar as dificuldades encontradas ao longo desta caminhada;

ao Departamento de Investigação Científica e Formação Avançada do Ministério do Ensino Superior, ao Instituto de Gestão de Bolsas de Estudo, assim como à Universidade Katyavala Bwila, em particular ao seu Instituto Superior Politécnico, por terem-me dado esta oportunidade e terem confiado na minha capacidade para chegar até aqui;

à Universidade da Beira Interior, o seu corpo docente e não docente por ter-me acolhido durante todo o período da minha formação;

ao meu orientador, o Prof.  Doutor João Paulo Cordeiro, pela sua dedicação, comprensão, apoio, e incentivo na realização desta dissertação;

à minha família, pelo amor, apoio e incentivo, sobre tudo nos momentos mais difíceis;

Em fim a todos que direta ou indiretamente deram os seus contributos na minha formação.

A todos, o meu muito obrigado.

# Resumo

A expansão da Internet e o surgimento das redes sociais suscitam o constante crescimento de texto on-line. Para ajudar as organizações no controlo do ponto de vista dos seus clientes emitido nesse novo canal de comunicação, surgiu a Análise de Sentimento em Texto (AST). Esta ciência ocupa-se no desenvolvimento de sistemas informáticos para previsão de sentimento em grande quantidade de texto não estruturado.

Por ser uma área recente, este subdomínio do Processamento de Linguagem Natural sofre de carência de recursos para o texto do domínio económico em Português. Face a esta realidade, com a preocupação de dotar o Português, a semelhança do Inglês, de conhecimento, ferramentas e recursos para AST no domínio económico, neste trabalho, verificou-se se os léxicos genéricos de AST em Português apresentam bons resultados quando são utilizados em domínio específico. Trabalhou-se particularmente com o domínio económico.

Para tal, desenvolvemos o SentiSoft, sistema de AST. Utilizando bases de dados lexicais como Sentilex-Pt e OpLexicon, a taxa de acerto variou, nas experiências em texto genérico, entre 81% e 74%, portanto, em texto do domínio económico, abaixo do 35%. A variação do sentido semântico dos vocábulos em função do contexto foi apontado como principal causa deste insucesso. Deste modo, concluí-se que os léxicos genéricos em Português, não apresentam bons resultados quando são utilizados em domínios específicos e sugeriu-se a elaboração de um léxico exclusivo para o domínio económico.

# Palavras-chave

Análise de sentimento em Texto, léxico, domínio económico, taxa de acerto.

# Abstract

The expansion of the Internet and the emergence of social networks are provoking the constant growth of online text. In order to help organizations in the control of their customers' opinions, issued in this new communication channel, arised the Text Sentiment Analysis (TSA). This science is concerned with the development of computer systems to classify sentiment in large amounts of unstructured text.

Being a recent area, this subdomain of Natural Language Processing suffers from lack of resources for the text of the economic domain in Portuguese. Given this reality, with the aim of equipping Portuguese, the similarity of English, knowledge, tools and resources for AST in the economic domain, in this work, it was verified that the generic AST lexicons in Portuguese present good results when they are used specific domain. They worked particularly with the economic domain.

For this, we developed the SentiSoft, AST system. Using lexical databases such as Sentilex-Pt and OpLexicon, the hit rate ranged from 81 % to 74 %, hence in economic domain text, below 35 %. The variation of the semantic sense of the words in relation to the context was pointed out as the main cause of this failure. Thus, it was concluded that the generic lexicons in Portuguese do not present good results when they are used in specific domains and it was suggested the elaboration of a lexicon exclusively for the economic domain.

# Keywords

Text Sentiment Analysis, lexicon, economic domain, hit rate.

# Conteúdo

# Lista de Figuras

# Lista de Tabelas

# Lista de Acrónimos

| | |
|---|---|
| ANEW | Affetive Norm for English Word |
| AST | Análise de Sentimento em Texto |
| DF | Document Frequency |
| DT | Decision Tree |
| FCE | Frequently Co-occurring Entropy |
| IMDb | Internet Movie Data base |
| LIWIC | Linguistic Inquiry and Word Count |
| PLN | Processamento de Linguagem Natural |
| SVM | Support Vector Machine |
| TM | Text Mining |

# Capítulo 1

# Introdução

O presente trabalho enquadra-se na área de Processamento de Linguagem Natural, mais concretamente em Análise de Sentimento em Texto. Neste primeiro capítulo, é apresentado o assunto a abordar nesta dissertação. O mesmo começa por uma descrição geral, seguida dos objetivos, motivação, exemplos e termina com a estrutura do trabalho.

## 1.1 Descrição geral

Nas organizações, obter informações certas em momento certo, é fundamental para avaliação das atividades realizadas, a fim de oferecer ao mercado os produtos e serviços equivalentes às exigências dos consumidores. Neste processo, a comunicação entre empresas e consumidores finais desempenha um papel determinante. Para se manter a evolução no mercado, os gestores das empresas preocupam-se em saber o nível de satisfação dos seus clientes quanto à qualidade de produtos consumidos e serviços beneficiados. Tradicionalmente, algumas das formas mais utilizadas para obter essa informação é a organização periódica de sondagens de opiniões de consumidores, através de chamadas telefónicas, e-mails, preenchimento de livros de reclamações e de questionários, até mesmo da realização de entrevistas aos clientes.

Nos últimos anos, o surgimento das redes sociais, a evolução e a proliferação de equipamentos de telecomunicações, como computadores, smartphones e outros meios, transformaram significativamente a forma como as pessoas transmitem as suas opiniões sobre diferentes assuntos. Milhares de pessoas estão conectadas e a trocarem informações, através da Internet, o tempo todo. Em Portugal, mais de metade dos internautas visitam as redes sociais várias vezes ao dia; o uso de conversa on-line (chat) e o comentar as publicações são os serviços mais utilizados [Reg16]. Como consequência desta revolução tecnológica e de comportamento, grandes volumes de texto formam-se e crescem a cada instante. Certamente, a maior fração destes volumes de informação é carregada de opiniões, tanto positivas como negativas.

O processo de tomada de decisão das pessoas é afetado pelas opiniões formadas por líderes de pensamento e pessoas comuns [Fel13]. Atualmente, quando uma pessoa quer comprar produtos ou contratar serviços, normalmente começa por pesquisar opiniões escritas por outras pessoas sobre as várias ofertas existentes. Esta mudança de comportamento dos clientes trouxe para as organizações novas oportunidades e desafios significativos. Estas querem acompanhar o ritmo do mercado (cada vez mais competitivo) e aproveitar o máximo possível as vantagens desta revolução tecnológica. Por isso, introduziram, no exercício das suas atividades, novas estratégias de trabalho: o comércio eletrónico, publicidades de marcas e produtos, etc. Apesar dessas vantagens, esta inovação tem as suas partes negativas, como a extinção do sigilo nas opiniões dos consumidores. Esta informação que, durante séculos foi restrita às organizações, tornou-se pública. Desta forma, a sua gestão e fiscalização em tempo real é um fator determinante para sobrevivência e o crescimento das empresas. Esta informação pode ser aproveitada pelas

empresas para conhecer as suas principais falhas. Com isso, será possível apresentar soluções adequadas às oscilações de procura e oferta do mercado.

Entretanto, perante o constante crescimento de texto on-line (informação não estruturado), extrair manualmente o conhecimento ali contido, torna-se uma tarefa muito complexa. Nesta situação, a melhor alternativa é o recurso às ferramentas que buscam automaticamente informação relevante em grande quantidade de texto. Perante esta situação e para dar solução a este problema, surgiu, dentro do Processamento da Linguagem Natural, uma nova área designada de *Análise de Sentimento em Texto* (AST). Esta área visa analisar opiniões, sentimentos, avaliações, atitudes e emoções das pessoas em relação a produtos, serviços, organizações, indivíduos, questões, eventos, tópicos e os seus atributos [Liu07]. Utilizando esta nova forma de extração de opiniões, as organizações poderão usufruir de inúmeras vantagens em relação à tradicional sondagem de opiniões de clientes (questionários, entrevistas, etc.), tais como:

a) A diminuição considerável da quantidade de documentos físicos, bem como o tempo e os gastos relacionados à sua utilização;

b) A obtenção de resultados de forma mais ágil e em tempo real;´

c) A extensão para um conjunto de estudo mais amplo e diversificado.

## 1.2 Objetivo e tarefas de investigação

A Análise de Sentimento em Texto (AST) permite identificar a opinião de pessoas em grandes volumes de dados não estruturados, cuja leitura e análise humana é inviável [Liu07]. Utilizando várias técnicas, como as de *Text Mining*, ela cria uma base de conhecimento contendo opiniões polarizadas (positivo, negativo ou neutro) [Liu07]. Esse processo começa desde a recolha de dados, o seu processamento até a previsão de sentimento. Atualmente existem vários métodos disponíveis para realização do processo de AST. Dentre eles, os mais utilizados são: os métodos baseados em léxicos e os sustentados por técnicas de Aprendizagem Automática. No entanto, existe escassez de recursos para textos em língua portuguesa. A maior parte desses recursos são restritos ao Inglês. Além disso, um grande número de dicionários disponíveis são genéricos, isto é, não são específicos a uma determinada área de estudo. [KB13] e [Bel17] defendem a teoria de que, os léxicos específicos a um domínio apresentam os melhores resultados na tarefa de classificação. Da mesma forma, [ST08] sustenta o conceito de que os classificadores de sentimento são fortemente dependentes de domínio. Portanto, a área económica não faz exceção a esta regra.

Deste modo, para responder às indagações acima apresentadas, neste trabalho pretende-se, verificar se um léxico específico para o domínio económico poderá trazer vantagens na previsão automática de valores sentimentais. Para alcançar este objetivo, foram traçadas as seguintes tarefas:

a) Estudo da área de Análise de Sentimento em Texto;

b) Análise linguística.

c) Análise linguístico-computacional;

d) Análise computacional

e) Experiências básicas e avançadas;

f) Análise dos resultados.

## 1.3 Motivação

O uso das ferramentas tecnológicas de informação pode mudar significativamente o perfil de uma empresa. Atualmente, isso desempenha um papel importante no crescimento das organizações, trazendo mais agilidade, eficiência, qualidade, entre outras vantagens na realização das suas atividades [Sua00]. Hoje é muito difícil falar de inovação, mudança, transformação, organização sem que em algumas partes desse processo a Informática não esteja envolvida. A História ensina-nos que as organizações começaram a crescer a partir da revolução industrial, quando se inicia a primeira revolução tecnológica, saindo de artesanal para o industrial. Sobre as indústrias, até ao século passado, falava-se de capitais avaliados em milhões de dólares, quando entram as tecnologias de informação, as organizações começaram a crescer de uma forma exponencial, começou-se a ter empresas globais, transcontinentais; hoje existem empresas bilionárias. Tal facto só é possível se as empresas tiverem ferramentas tecnológicas completas que lhes dão o mínimo de trabalho possível na execução das suas tarefas. Ao nível de países lusófonos, quanto à análise automática de sentimento em texto, as ferramentas utilizadas não trazem ainda grande satisfação aos seus utilizadores, pelo facto de a maior parte delas estarem desenvolvidas exclusivamente para texto em Inglês. Na maior parte dos trabalhos com o texto em Português, primeiro faz-se a tradução para Inglês, depois a sua classificação. A dificuldade torna-se maior quando se trata de texto restrito a uma determinada área de estudo.

Nos últimos anos, os resultados do interesse de investigadores em AST têm sido visíveis. Vários recursos foram lançados no mercado, como aplicações para recolha automática de comentários de utilizadores de redes sociais, os dicionários de palavras de sentimento, os lematizadores, sistemas de aprendizagem automática, entre outros. Por ser uma área nova e ainda em desenvolvimento, apesar de o Português ser umas das línguas mais faladas do mundo, estas ferramentas não estão projetadas para esta língua. Para mitigar as necessidades dos utilizadores lusófonos, geralmente têm sido feito adaptações a partir de recursos desenvolvidos para a língua inglesa. O problema é o seguinte: nesse processo as especificidades de cada língua não são consideradas, o que acarreta algumas consequências nefastas quanto à qualidade dos resultados obtidos.

Os problemas com a qualidade da informação são sentidos de forma rotineira por todas as organizações, com diferentes níveis de gravidade e de prejuízo [JNO99]. O impacto negativo traduz-se em custos desnecessários, em processos de decisão afetados ou na perda de confiança dos clientes [JNO99]. Deste modo, a escolha deste tema justifica-se pela necessidade de dotar o Português, à semelhança do Inglês, de técnicas, recursos e conhecimento de AST no domínio económico. Assim, propõe-se o desenvolvimento de um sistema, no qual serão realizadas experiências com vários conjuntos de frases da área económica, utilizando bases de dados lexicais como SentiLex-PT [CP15] e OpLexicon. A análise dos resultados dessas experiências poderá trazer luz à seguinte questão: **Será que os léxicos genéricos em Português para Análise de Sentimento em Texto apresentam bons resultados quando são utilizados na classificação de texto do domínio económico?**

## 1.4 Exemplo

O texto é o recurso primordial para exposição de opiniões. Este pode exprimir um sentimento de satisfação (positivo), de insatisfação (negativo) ou nenhum dos dois (neutro). Nesta secção, ilustrámos, a partir de exemplos, os casos relacionados a estes três sentimentos. A Tabela 1.1 apresenta algumas palavras com os seus respetivos sentimentos:

| **Termo** | **Sentimento** |
|---|---|
| bom | Positivo |
| inteligência | Positivo |
| maldição | Negativo |
| ausência | Negativo |
| casa | Neutro |
| elogio | Positivo |
| ferramenta | Neutro |

Tabela 1.1: Palavras e respetivos sentimentos.

O sentimento de uma frase, geralmente, depende das palavras que a constituem. A Tabela 1.2 apresenta as classificações de algumas frases.

| **Categoria** | **Sentimento** | **Frase** |
|---|---|---|
| **Twitter ou SMS** | **Positivo** | Estou muito feliz, é o dia do meu aniversário…. |
| | **Negativo** | Não aguento mais, fim da relação. |
| | **Neutro** | Ok te vejo na segunda, trarei o livro!!!!!! |
| **Jornal** | **Positivo** | Os professores estão satisfeitos com o acordo entre o sindicato e o governo, as aulas retomam o seu curso normal a partir de segunda-feira (16). |
| | **Negativo** | O A filha do chefe do estado-maior foi raptada e violada na madrugada do sábado. |
| | **Neutro** | O ministro do interior, na sua vinda à província mais ao sul do país, manteve um encontro com as autoridades locais. |

Tabela 1.2: Frases e repetivos sentimentos.

Como podemos notar, nos Jornais, as frases têm uma estrutura formal com palavras escritas corretamente, enquanto que no Twitter ou SMS, o texto, geralmente, é informal, com liberdade e imprecisão.

## 1.5 Estrutura do trabalho

Este trabalho está planeado e organizado segundo a seguinte estrutura:

a) Capítulo 1: Introdução.

b) Capitulo 2: Enquadramento e Estado da arte - Fundamenta teoricamente os aspetos ligados à Análise de Sentimento em Texto e às áreas relacionadas, assim como os trabalhos relacionados a esse tema de investigação.

c) Capítulo 3: Caracterização do sistema - Trata de padronizar e codificar as informações adotadas para realização do processo de Análise de Sentimento em Texto.

d) Capítulo 4: Experiências e Resultados - Realiza as experiências com texto genérico e do domínio económico, em função dos resultados, apresenta as contribuições;

e) Capítulo 5: Conclusões e Trabalho futuro.

# Capítulo 2

# Enquadramento e Estado da Arte

O presente capítulo trata de relatar todos os aspetos teóricos de Análise de Sentimento em Texto e áreas relacionadas. Para a sua melhor organização, está dividido em três secções: enquadramento, estado da arte e sumário.

## 2.1 Enquadramento

Esta secção relata os conceitos das áreas relacionadas a esta investigação e também descreve detalhadamente as metodologias e técnicas utilizadas no desenvolvimento deste trabalho.

### 2.1.1 Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) é uma área da Inteligência Artificial que permite a aplicação de teorias computacionais para compreender e responder naturalmente às questões relacionadas à linguagem humana, tornando possível o desenvolvimento de sistemas com capacidade de reconhecer e produzir informação apresentada em linguagem natural.

A investigação em PLN baseia-se em todos os aspetos essenciais da comunicação em linguagem natural: Fonologia, Morfologia, Sintática, Semântica e Pragmática [MAC95]. A Fonologia estuda o som produzido por palavras; a Morfologia determina as classes gramaticais de termos da frase; a Sintática relaciona-se com a função de vocábulos da oração; a Semântica ocupa-se do significado de palavras [MAC95]; finalmente, a Pragmática estuda a linguagem no contexto da sua utilização na comunicação. Como podemos notar, a Linguística é a companheira indispensável do PLN, para além dela, existem outras ciências como a Aprendizagem Automática e a Matemática, que o auxiliam na realização das suas tarefas. O PLN tem a sua aplicação em áreas como: Pesquisa de Informação, Extração de Informação, Sumarização Automática, Tradução Automática, Diálogo Homem-Máquina, etc. Esta dissertação baseia-se numa das áreas de aplicação de PLN mais promissora, a Análise de Sentimento em Texto.

### 2.1.2 Análise de Sentimento em Texto

A Internet e a proliferação de dispositivos móveis inteligentes mudaram a forma como as informações são criadas e divulgadas. No mundo atual, a maior parte das pessoas partilham pensamentos e opiniões sobre qualquer assunto através de redes sociais como o Facebook, Twitter e sistemas de mensagens instantâneas como o Skype e WhatsApp. Esse facto resultou na proliferação de conteúdo textual. Naturalmente, essa abundância de dados rapidamente atraiu interesse de negócios e pesquisas de vários campos, incluindo marketing, ciências políticas e sociais, etc. Todos estão interessados em ter respostas às perguntas como:

a) Será que as pessoas gostam do nosso novo produto?

b) O design do novo produto que acabamos de lançar está a agradar os consumidores?

c) O que os clientes dizem sobre a nossa forma de atendimento?

a) Será que os cidadãos apoiam o nosso programa?

b) Podemos vencer as próximas eleições?

Para dar respostas certa e em tempo real a essas e outras perguntas, surgiu a Análise de Sentimento em Texto (AST). Esta área de Processamento de Linguagem Natural especializou-se no desenvolvimento de sistemas capazes de detetar automaticamente sentimentos expressos em textos. Esse sentimento pode ser positivo, negativo ou neutro. Assim, em pouco espaço de tempo, as organizações podem facilmente acompanhar as tendências do mercado e obter feedback dos clientes sobre os seus produtos e serviços.

A AST preocupou-se com o desenvolvimento de métodos eficazes que fornecem informações suficientes a partir das quais poder-se-á determinar opiniões expressas em qualquer novo item (tweet, frase, título, trecho ou texto inteiro). A qualidade de um método de análise automática de sentimento é avaliada em função de aproximação dos seus resultados à avaliação humana. Para cumprir com esta tarefa, ao longo do tempo foram desenvolvidos vários métodos, repartidos em dois grandes grupos: os métodos baseados em Aprendizagem Automática e os baseados em léxico.

### 2.1.3 Métodos baseados em Aprendizagem Automática

A Aprendizagem Automática é a tecnologia que dá aos computadores a capacidade de aprender sem ser explicitamente programados [Dat18]. Este subconjunto da Inteligência Artificial permite habilitar os computadores a aprenderem como os seres humanos. A Aprendizagem Automática divide-se em dois grandes grupos: Aprendizagem Supervisionada e não Supervisionada.

A Aprendizagem Supervisionada consiste em treinar o computador através de um conjunto de exemplos compostos de dados de entrada e as suas respetivas soluções, de maneira a permitir que os seus algoritmos criem conceitos gerais a partir das características comuns identificadas nos exemplos. Estes servir-lhe-ão de base para a resolução de novos problemas da mesma natureza sem ajuda humana.

A Aprendizagem Supervisionada é muito utilizada em sistema de reconhecimento facial. Por exemplo, para ensinar ao computador a reconhecer uma determinada pessoa chamada $X$, deve-se lhe apresentar várias fotos do $X$, assim, os algoritmos de Aprendizagem Automática vão fazer coleta das características do rosto dele até criar um padrão. Deste modo, nas próximas vezes que se apresentar uma nova foto, o computador fará uma comparação das características encontradas nesta foto às do modelo padrão. Se forem iguais, conclui que se trata de $X$, senão, a conclusão será que se trata de outra pessoa.

Na Aprendizagem não Supervisionada, a partir de um conjunto diversificado de dados de entrada, o sistema tira as suas próprias conclusões. Vamos manter o exemplo do parágrafo anterior, supondo que agora temos várias fotos não só de $X$, mas também de outra(s) pessoa(s). Com isso, queremos que essas imagens sejam categorizadas. O computador vai analisar cada uma dessas fotos, criando um modelo de características para cada uma delas e agrupar os dados mais parecidos (aglomeração).

O grande problema dos algoritmos de Aprendizagem Supervisionada é que são dependentes do domínio de treino. Isto é, apresentam bons resultados quando são treinados e testados em dados do mesmo domínio e o seu desempenho pode baixar significativamente quando o teste é feito em dados de domínios diferentes [ST08]. Mesmo conhecendo antecipadamente os padrões, alguns programadores preferem utilizar a Aprendizagem não Supervisionada ao invés da supervisionada, pois, definindo os padrões, limita-se a busca dos algoritmos de Aprendizagem Automática; eles não vão procurar por novas características que talvez nem o ser humano tinha notado [CL09].

A Análise de Sentimento em Texto recorre às técnicas de Aprendizagem Automática para a realização de várias operações, como, por exemplo: análise sintática, análise morfológica, previsão de valores sentimentais, etc.

Existem vários algoritmos de aprendizagem automática como: Naive Bayes, K Nearest Neighbors (KNN), Redes Neuronais (ANN), Auto Regressive Moving-Average (ARMA), Support Vector Machines (SVM), etc.

### 2.1.4 Métodos baseados em léxico

De acordo com a Linguística, o léxico é um conjunto de palavras vinculadas a uma língua, seguidas de dados sintáticos, semânticos e morfológicos. [YW00] chama o léxico de dicionário tratável por máquina, MTD (Machine-Treatable Dictionary), por ter formato lexical que facilita o tratamento da sua informação por sistema de PLN. O mesmo autor defende que existe uma diferença entre o MTD e o dicionário legível por máquina, MRD (Machine Readable Dictionary). Este último é um dicionário comum em formato digital, criado por lexicologistas, apenas para o uso humano; enquanto o MTD é concebido especificamente para ser utilizado em sistemas informáticos. Para além de dados linguísticos, cada palavra do léxico contém o seu valor sentimental (positivo, negativo ou neutro). Este é expresso através de um valor numérico denominado polaridade. Muitos léxicos acrescentam ao valor sentimental, informações sobre a força, isto é, quão positivo ou negativo é o termo. Se é fortemente positivo, moderamente positivo, ligeiramente positivo, fortemente negativo, moderadamente negativo, ligeiramente negativo. A polarização pode ser feita manualmente ou através das técnicas da Aprendizagem Automática.

Ao longo do tempo, vários léxicos foram elaborados, a maior parte deles para a língua inglesa. Dentre eles, encontram-se os seguintes: SentiWordnet, Affective Norms for English Words (ANEW), Afinn, Hapiness Index, Oplexicon e SentiLex-Pt. Estes dois últimos merecem destaque por serem uns dos raros dicionários de palavras com sentimento criados exclusivamente para a língua portuguesa e por serem também os recursos usados nesta investigação.

O **Sentilex-Pt** é repartido a dois léxicos associados: o de lemas e o de formas flexionadas [CP15]. O primeiro contém 7 014 lemas, repartidas em 4 770 adjetivos, 1 081 nomes, 666 expressões idomáticas e 489 verbos. O segundo é composto de 82 347 formas flexionadas, associadas aos respetivos lemas e repartidas em: 16 863 adjetivos, 1 280 substantivos, 29 504 verbos [Bel17]. Para evitar ambiguidade na sua polarização, as entradas de Sentilex-pt são restrintas às entidades humanas. Os valores de polaridade neste léxico são 1 (positivo), 0 (negativo) e -1 (negativo).

```
bonito.PoS=Adj;TG=HUM:N0;POL:N0=1;ANOT=MAN
castigado;PoS=Adj;TG=HUM:N0;POL:N0=-1;ANOT=JALC
estimado.PoS=Adj;TG=HUM:N0;POL:N0=1;ANOT=JALC;REV=AMB
enganar.PoS=V;TG=HUM:N0:N1;POL:N0=-1;POL:N1=0;ANOT=MAN
```

Figura 2.1: Extrato de Sentilex-Pt ( Fonte: [CP15]).

O **OpLexicon** (Opinion Lexicon) é um lexico de sentimento para língua portuguesa. Este recurso contém 32 653 entradas repartidas em 32 144 palavras, 462 expressões idiomáticas e 47 emoticons.

### 2.1.5 Procedimento metodológico

O desenvolvimento de sistemas, capazes de processar informações subjetivas de maneira eficaz, requer a superação de uma série de desafios. Para a sua melhor organização, esta dissertação baseou-se na metodologia proposta em [DdS05], na qual o trabalho de Processamento da Linguagem Natural é repartido em três domínios de investigação:

a) domínio linguístico;

b) domínio representacional ou linguístico-computacional;

c) domínio computacional.

No domínio linguístico foram tratados os aspetos relacionados com o Português como análise sintática, semântica e morfológica; no domínio representacional, as informações gerais selecionadas e organizadas no domínio anterior foram moldadas numa representação tratável pelo computador; por fim, no domínio computacional, as representações computáveis foram utilizadas para desenvolver um sistema informático para Análise de Sentimento em Texto.

### 2.1.6 Recolha de dados

A matéria-prima para a realização do processo de Análise de Sentimento em Texto consistiu em dados. Nesse caso trata-se de dados textuais que geralmente têm como proveniência as redes sociais. O Processamento da Linguagem Natural oferece-nos várias ferramentas para recolha de grande quantidade de dados na internet, dentre elas, salientamos: o Discovertext, o Rapidminer e o SAS Web Crawler [Oli15]. Alguns sites de microblog (como Facebook, o Twitter e o Sina-Wiebo) disponibilizam as suas API (Application Programming Interface) para aquisição dos seus dados públicos. O Facebook fornece o Facebook Graph API4 e a Sina-Wiebo, o Tancent API5 [AO13]. Essas aplicações permitem extrair posts e outras informações relacionadas a esses sites web. Da mesma forma o Twitter disponibiliza uma API REST para obtenção de dados estatísticos sobre os seus utilizadores e a Streaming API2 para obter dados de streaming - os tweets [KML13].

### 2.1.7 Pré-processamento

O conteúdo de texto a analisar, sobretudo quando provém da web, apesar da riqueza que possui em termo de opiniões, aparece na forma de texto livre, isto é, com abreviaturas, números, erros ortográficos, etc. De uma forma geral podemos dizer que o conteúdo desse texto não possui padrões necessários para a sua análise, por isso, é chamado de texto não estruturado. Em Petroquímica, pode ser comparado ao petróleo no seu estado bruto, deve ser refinado para

se obter os seus derivados (combustível, Asfalto, Lubrificantes, etc.). Por isso, é necessário fazer primeiro, uma revisão cuidada de cada pedaço de texto para permitir a sua organização e melhorar o seu uso nas fases subsequentes de análise.

Para a realização do processo de estruturação de dados, são utilizadas as técnicas de Text Mining (TM). Considerada uma evolução da área de Recuperação de Informações (RI), o Text Mining é um processo de descoberta de conhecimento que utiliza as técnicas de análise e extração de dados a partir de textos, frases ou apenas palavras [EAMM07]. Este método de recuperação de informações é aplicada para o processamento de grandes quantidade de dados e envolve várias tarefas (técnicas). Neste trabalho, de acordo com o objetivo a alcançar, resumimos o pré-processamento nas seguintes tarefas de TM: Tokenização e Normalização.

A **Tokenização** é a operação que permite transformar o texto num conjunto de termos isolados chamados de Tokens. Geralmente, os tokens extraídos em texto são palavras, sendo facilmente identificados quando se utiliza o espaço como delimitador.

A **Normalização** é uma operação complexa composta por várias etapas que apresentamos nos pontos a seguir:

a) **Correção ortográfica**: O objetivo desta etapa é restituir cada palavra na sua forma normal de escrita, por isso, os erros ortográficos são corrigidos, da mesma forma, as abreviaturas e siglas são reescritas com palavras completas, os ícones de emoções são substituídos pelas suas correspondentes palavras sinónimas. Existem casos em que as palavras estão mal escritas gramaticalmente com o objetivo de intensificar o sentimento nelas expresso. Para esta situação, para que essas palavras não percam o sentido pelo qual foram utilizadas, devem ser criadas regras específicas para o seu tratamento. Como, por exemplo, os internautas têm o hábito de escrever *boooom* que significa *muito bom*;

b) **Transformação de letras maiúsculas em minúsculas**: Nesta etapa, transforma-se todas as letras maiúsculas em minúsculas para permitir uniformizar a escrita;

c) **Supressão de stop words**: Nesta etapa, elimina-se todas as palavras consideradas irrelevantes. Como por exemplo, os artigos, as preposições, os pronomes, etc. São eliminados também as marcações (HTML), os espaços em branco, os acentos e qualquer símbolo que não seja alfanumérico;

d) **Supressão de variações morfológicas**: Durante esta etapa são realizadas as seguintes operações: Lematização e Stemming. A primeira operação transforma uma palavra na sua forma canónica; e a segunda visa encontrar a raiz morfológica de um termo, eliminando os seus prefixos e sufixos. A lematização é aplicada em verbos conjugados. Por exemplo, os termos *comerei, como, comeis, comi*; lematizados, todos são transformados num único termo, *comer*. A eliminação de variações morfológicas necessitam de grande conhecimento do vocabulário e análise completa da língua [Kau16]. Ela visa aumentar o desempenho de sistema de PLN. [AGLdS00] alega que, em análise de emoções, a lematização não tem influência significativa no resultado.

Para ilustração, considera-se o seguinte texto como exemplo:

*A partir da próxima semana, as aulas do subsistema do ensino não universitário vão paralisar em todo o país durante três semanas, anunciou o Sindicato Nacional dos Professores de Angola, que pondera repetir a greve também em junho.*

Após o seu pré-processamento, ter-se-á:

*próxima semana aula subsistema ensino não universitário ir paralisar pais durante semanas anunciar sindicato nacional professor angola ponderar repetir greve também junho*

Nota-se que após o pré-processamento, a extensão do texto é reduzida.

### 2.1.8 Etiquetagem morfológica

Nesta fase determina-se a classe gramatical de cada palavra da frase, pois, tem grande importância no processo de Análise de Sentimento em Texto. De acordo com [VH00], os adjetivos são fortes indicadores de subjetividade. Na maior parte dos léxicos de sentimento, a cada palavra, para além da sua polaridade, é-lhe associada também uma classe gramatical, isto é, se é um substantivo, um verbo, um adverbio, etc.

### 2.1.9 Previsão de valores sentimentais

Esta é a etapa da Análise de Sentimento em Texto propriamente dita, pois, é aqui onde é determinado o valor sentimental escondido numa determinada opinião. Uma opinião é um par ordenado composto por dois elementos fundamentais: o alvo e o valor sentimental. O alvo é a entidade sobre qual é emitida a opinião e o valor sentimental ou polaridade corresponde a quantidade de sentimentos expressos.

A previsão de sentimento pode ser feita nos três níveis seguintes: nível da palavra, da frase e do texto. Os parágrafos a seguir descrevem esse processo.

Numa primeira fase, o texto a ser analisado é separado em frases. A seguir, é determinado o valor sentimental ao nível da palavra através de busca direta, no léxico, da polaridade correspondente a cada termo da frase. Dependendo do sinal da polaridade, o resultado pode ser positivo, negativo ou neutro. No caso concreto dos léxicos Sentilex-PT e OpLexicon, o valor para cada palavra das frases é -1,1 ou 0. Depois é calculado o valor sentimental da frase. Sobre este assunto, existem várias abordagens que se diferenciam nas fórmulas utilizadas. O ponto de convergência entre elas é que, na realização desta tarefa, todas consideram a polaridade das palavras que compõem a frase.

[Nie14] prevê o sentimento da frase pela relação entre as polaridades das palavras que a constituem e o seu comprimento. Deste modo, calcula o seu valor sentimental pelo quociente da soma de polaridades das suas palavras e a raiz quadrada do seu comprimento.

$$Polaridade(frase) = \frac{1}{\sqrt{\mid frase \mid}} * \sum_{w \in frase} Polaridade(w) \quad \textbf{(2.1)}$$

Por exemplo, para a frase: *Os alunos inteligentes estão atentos*, a polaridade segundo a fórmula acima referênciada é:

$$Polaridade(frase) = \frac{1}{\sqrt{5}} * (0 + 0 + 1 + 0 + 1) = \frac{2}{\sqrt{5}} = 0.89 \quad \textbf{(2.2)}$$

A atribuição de polaridade às palavras, neste exemplo, teve como base o Sentilex-Pt, no qual *inteligente* e *atento* tem polaridade $1$.

[AE06] propõe um método que estima o valor sentimental da frase através da soma de polaridades de todos os seus termos.

$$Polaridade(frase) = \sum_{w \in frase} Polaridade(w) \quad \textbf{(2.3)}$$

Aplicando este conceito nas frases abaixo, temos:

a) *Estamos a ser empurrados para a greve, referiu também o responsável. estar (0) ser (0) empurrado (-1) greve (-1) referir (-1) responsável (0)*

$$Polaridade(frase1) = 0 + 0 + (-1) + (-1) = -2 \quad \textbf{(2.4)}$$

b) *A nossa seleção é a melhor do continente. seleção (0) ser (0) melhor (1) continente (0)*

$$Polaride(frase2) = 0 + 0 + 1 + 0 = 1 \quad \textbf{(2.5)}$$

c) *O cão ladrou uma vez nesta noite. cão (0) ladrar (0) noite (0)*

$$Polaridade(frase3) = 0 + 0 + 0 = 0 \quad \textbf{(2.6)}$$

No primeiro exemplo, a opinião expressa na frase é negativa; no segundo, o sentimento da frase é positivo; e no terceiro caso, a polaridade é zero.

Quanto ao nível do texto, [Tab16] defende que a previsão de sentimento deve ser feita através da soma de polaridades de todas as suas frases.

$$Polaridade(texto) = \sum_{frase \in texto} Polaridade(frase) \quad \textbf{(2.7)}$$

Se, por exemplo, juntarmos as frases dos três exemplos anteriores para formar um texto, teremos: *Estamos a ser empurrados para a greve, referiu também o responsável. A nossa seleção é a melhor do continente. O cão ladrou toda a noite.*
Neste caso, somando as polaridades das três frases, temos o valor sentimental do texto.

$$Polaridade(texto) = -2 + 1 + 0 = -1. \quad \textbf{(2.8)}$$

No entanto, nem todos concordam com esta abordagem. Se um texto for rico em opiniões e as suas frases tiverem polaridades opostas, aplicando a abordagem acima proposta, corre-se o risco de se obter como resultado o valor 0, que corresponde ao sentimento neutro e que nesta situação, não reflete bem a verdade. Por esta razão, existem outras abordagens para este nível

de AST, nas quais algumas propõem a polaridade mínima ou máxima das frases do texto e outras a média dos dois valores.

Assim sendo, procedendo de uma ou outras formas, é possível prever a polaridade das frases e textos, porém, essas formas de determinar o valor sentimental, em algumas situações, não produzem resultados satisfatórios, dada a ambiguidade e complexidade da língua. Nos passos a seguir, são apresentados os casos que se relacionam a estas situações, bem como as propostas de soluções para ultrapassar os constrangimentos que podem causar no processo de AST.

### 2.1.10 Expressões idiomáticas

Uma expressão idiomática é um conjunto de duas ou mais palavras que se caracterizam por não ser possível identificar o seu significado mediante o sentido literal dos termos que a constituem [Wik18]. O seu significado não pode ser deduzido a partir dos significados das palavras que a constituem. Desta forma, tratar os elementos constituintes das expressões idiomáticas de forma independente faz perder o verdadeiro sentido da mensagem ali transmitida. Como consequência, aumenta a probabilidade de mudança do valor sentimental ali expresso. Analisemos o caso dos exemplos abaixo, vamos basear-nos no léxico Sentilex-Pt para a previsão de sentimento das frases ali apresentadas. Se tratarmos os termos constituintes de forma independente, a polaridade de cada frase será 0, mas o certo é considera-las como expressões idiomáticas. Neste último caso, o resultado será $1$ para a primeira, e $-1$ para a segunda:

a) *Ser um homem às direitas.*

*b) Andar a dormir à sombra da bananeira.*

A maior parte dos léxicos, nas suas entradas, para além de palavras, contempla também uma lista de expressões idiomáticas.

### 2.1.11 Negação

Em Análise de Sentimento em Texto, a negação merece uma atenção particular, pois, é uma operação que pode atuar ao nível sintático, semântico e pragmático da frase. Logo a sua presença tem grande influência no valor sentimental da frase que a contém. Aparentemente as frases: *Gosto deste livro* e *Não gosto deste livro* são muito semelhantes, mas a presença de um único token diferente, o termo de negação *não*, faz com que as duas frases tenham polaridades totalmente opostas.

Geralmente a presença da negação num texto é detetada através de advérbios de negação. Segundo [ An06b], *não* era considerado tradicionalmente como o único advérbio de negação, mas atualmente a gramática portuguesa já admite outros, como: *tampouco, nem, nunca, jamais*, etc.

Numa frase, o advérbio de negação pode negar a frase completa (*O Maurício não entregou dinheiro à Joana*), ou simplesmente um único elemento (*O António comprou à Gisela ontem, não cadernos, esferográficas*).

Nem sempre a negação é introduzida através de advérbios de negação. As locuções adverbiais de negação (*de modo algum, de jeito nenhum, de forma nenhuma, etc.*), as conjunções adversativas (*mas, porém, etc.*) e alguns pronomes indefinidos (*nenhum, ninguém, nada, nunca, etc.*) também podem desempenhar esse papel. Os exemplos a seguir ilustram essas situações:

a) *O aluno começou bem, mas acabou por ficar pelo caminho.*

b) *Nesta prova, nenhum estudante teve positiva*.

| **Termos da negação** |
|---|
| não |
| nenhuma |
| de jeito nenhum |
| nada |
| nunca |
| jamais |
| Nem |

Tabela 2.1: Termos da negação.

Um termo da negação também pode ser empregue em outras expressões sem constituir uma negação da proposição expressa. Como exemplificado nas frases abaixo:

a) *Este telefone não é apenas caro, mas também pesado e difícil de usar.*

*b) Este estudante não só é inteligente, mas também dedicado.*

Como pode ser visto a partir dos exemplos acima, a modelagem da negação é um aspeto importante em Análise de Sentimento em Texto, porém, difícil e altamente variável. Ela é influenciada por vários fatores como o impacto que o autor quer fazer sobre o significado geral do texto, o contexto, o género textual, etc., entretanto, nesta abordagem o objetivo de AST é identificar corretamente o âmbito da negação. Isto é, determinar com precisão a parte da frase modificada pela presença deste fenómeno linguístico. A solução a este problema passa geralmente na criação de um modelo de negação. Este modelo, para ser eficaz na Análise de Sentimento em Texto, geralmente, requer o conhecimento de **expressões polares** [MWDK10]. As Expressões polares são palavras que contêm polaridades prévias. Existem várias abordagens sobre este assunto em PLN. Algumas delas são as seguintes:

[PL06] propõe um modelo de negação, no qual as pontuações são atribuídas às expressões polares, ou seja, pontuações positivas às expressões polares positivas e negativas às expressões polares negativas, respetivamente. Se uma expressão polar é negada, a sua polaridade é simplesmente invertida. O exemplo a baixo ilustra este modelo.

a) *inteligente* $(+1) \rightarrow$ *não inteligente* $(-1)$

[JL09] sugere uma abordagem que considera uma expressão polar negada, como uma expressão polar não negada com polaridade oposta.
exemplo:

a) *A cerimónia terminou sem elogio.*

A palavra *elogio* (do exemplo acima), em Sentilex-Pt, tem polaridade $1$, [JL09] propõe-nos acrescer uma nova expressão polar, neste caso sem-elogio, com polaridade $-1$.

### 2.1.12 Amplificadores e atenuadores

A presença dos advérbios de intensidade nas frases merece também uma atenção especial, pois, estes termos têm grande influência nos sentimentos ali expressos. Nesta subsecção, vamos fazer uma análise desta situação. Os tradicionais advérbios de intensidade utilizados são: *bastante, muito, pouco, quase, tanto, tão* e *demais* [ An06a]. Esses termos, numa frase, funcionam como quantificadores universais, que aumentam ou diminuem a força polar dos adjetivos que modificam. A Tabela 2.2 mostra as duas classificações dos advérbios de intensidade.

| **Amplificadores** | **Atenuadores** |
|---|---|
| muito | pouco |
| demais | quase |
| bastante | tanto |
| bem | menos |
| tanto | apenas |
| deveras | menos |

Tabela 2.2: Atenuadores e amplificadores.

Os exemplos a seguir ilustram alguns casos relacionados a esta situação:

a) *Faça menos esforço.*

b) *João é muito inteligente.*

c) *João é pouco inteligente.*

A presença do adverbio *muito* à frente de *inteligente*, no exemplo da alinea b) aumenta o sentimento expresso neste adjetivo, enquanto na alinea c), o adverbio *pouco* reduz o sentimento expresso através do adjetivo *inteligente* , o mesmo acontece na alinea a).

Os modelos de negação sugeridos por [PL06] e [JL09], relatados na subsecção anterior, são extensíveis para os atenuadores e amplificadores. A única diferença consiste no valor da polaridade dos adjetivos modificados, que para este caso, não é invertido; só é alterada a intensidade. Como, por exemplo: *menos atraente* e *atraente* terão a mesma polaridade, mas intensidades diferentes.

### 2.1.13 Ironia

A ironia é um processo retórico de usar intencionalmente palavras ou expressões para proferir um significado diferente do que elas têm quando utilizadas literalmente [PC09]. No exemplo abaixo, o sentimento expresso é totalmente negativo, mas aparentemente, baseando-se nos termos ali empregues, a impressão é de uma opinião positiva.

a) *Como estiveste bem na prova, não tiraste nem a nota mínima!*

Se o aluno não teve a nota mínima, certamente a afirmação de que ele foi bem à prova é falsa. Os termos opostos, utilizados na segunda parte desta oração, produziram a ironia. Em vez de bem, a realidade é que o aluno foi muito mal na prova.

Devido à sua natureza, a ironia, geralmente, expressa uma opinião com sentimento. A ironia verbal pode ser expressa por figura de estilo como sarcasmo, hipérbole, metáfora, entre outras, cujas diferenças, na prática, pode ser bastante difícil de distinguir. [PC09], depois de analisar os erros do seu classificador, confirma a relevância da ironia em AST, afirmando que uma grande proporção de erros de classificação deriva da incapacidade do sistema em deteção de expressões irónicas. No mesmo trabalho, apresentam-se algumas indicações (como emoticons, expressões onomatopáicas para o riso, sinais de pontuação pesados, aspas e pontos positivos, interjeições) para detetar a ironia em comentários de internautas. Embora o uso dessas pistas nos padrões definidos ajuda a detetar algumas situações nas quais a ironia está presente, elas não representam completamente o fenómeno. Logo, a modelação da ironia continua a ser o grande desafio da Análise de Sentimento em Texto.

A partir das diferentes abordagens relatadas nesta secção sobre a previsão de sentimento em texto, podemos deduzir que a unidade básica de Análise de Sentimento em Texto é a palavra. A unidade fundamental é a frase, pois, é neste nível que são análizados e solucionados os problemas causados pelas situações relatadas.

## 2.2 Estado da arte

Nesta secção apresentamos um resumo sobre os trabalhos relacionados a esta investigação, pois esta informação é necessária para evitar perda de tempo com os trabalhos já realizados por outros. As publicações encontradas são repartidas em três grupos: o primeiro contém as abordagens baseadas em léxico; o segundo apresenta as baseadas em Aprendizagem Automática; e no último constam outros tipos de publicações que merecem destaque nesse trabalho.

### 2.2.1 Abordagens baseadas em léxico

Por carência de recursos em língua portuguesa para a classificação de emoções, [AGLdS00] realizou um estudo que visou analisar qual das duas abordagens seguintes produz o melhor resultado:

a) Tradução do texto a analisar para um idioma em que haja recursos para tratar de emoções (no caso o Inglês);

b) Tradução do léxico para língua na qual está escrito o texto.

Além disso, verificar se a aplicação da lematização contribui na melhoria do resultado. Depois de várias experiências, conclui-se que a tradução do texto da revisão apresenta melhor resultado que a tradução do dicionário; e quanto à lematização, esta prática não tem influência significativa no resultado.

A maioria dos trabalhos em AST apresenta uma avaliação global para o sentimento expresso no texto. Isto acontece pelo facto das palavras com sentimento assumirem polaridades pré-definidas. Essa forma de classificação não oferece ao utilizador final uma visão mais refinada das opiniões acerca da entidade avaliada. [NRS00] realizou uma análise mais profunda que permite considerar as características do objeto a ser avaliado. Sabendo que alguns adjetivos mudam de polaridade a depender do substantivo que eles qualificam (*cerveja quente, pizza quente*), o

processo consiste em identificar pares válidos (característica, palavra opinativa) que recebem uma polaridade única. Os Resultados experimentais provaram a eficácia do processo.

[SBS00] apresentou um recurso lexical, para apoiar na classificação de palavras com sentimento, chamado SentiWordNet 3.0 que é, na realidade, uma versão melhorada do léxico SentiWordNet 1.0. Este documento é o resultado de anotação automática de todos os sintagmas do dicionário WordNet de acordo com os seus graus de positividade, negatividade e neutralidade. Atualmente licenciados para mais de 300 grupos de pesquisa e utilizados em diversos projetos de investigação, esta ferramenta está assim disponível ao público.

[PM09] deu um contributo importante em Análise de Sentimento em Texto, pois, desenvolveu um recurso unificado que combina os conhecimentos léxicais e de Aprendizagem Supervisionada na categorização de texto. As experiências realizadas provaram que a combinação de informações lexicais com Aprendizagem Supervisionada produz melhores resultados que a utilização separada das duas ferramentas de previsão de sentimento.

[MT11] apresentou o SO-CAL (Semantic Orientation CALculator), novo dicionário de palavras anotadas com a sua orientação semântica (polaridade e força) para extrair sentimento em texto. Nesta nova ferramenta foram introduzidos intensificadores e refinada a abordagem à negação. Os resultados atuais são estatisticamente melhores em relação às versões anteriores do sistema SO-CAL.

[DT11] demostrou que a Análise de Sentimento em Texto pode ser utilizada para medir o nível de (in)sucesso que determinados produtos ou serviços poderão ter nos primeiros dias do seu lançamento. Para isso, utilizou as mensagens dos comentários de 10 novos filmes, 7 dias antes dos seus lançamentos no cinema para a indústria cinematográfica norte-americana. Foi observado que quanto mais positivo o filme fosse comentado, maior era venda do seu bilhete na sua semana de estreia.

O SenticNet é, atualmente, um dos mais abrangentes recursos semânticos disponíveis gratuitamente para a mineração de opiniões. Ele fornece apenas polaridades, mas não fornece informações mais detalhadas sobre os seus conceitos. Outro recurso importante para este efeito é o WordNet-Affect, que, por sua vez, é um simples dicionário de palavras, não possui informação quantantiva. [SP12] relatou um trabalho sobre a fusão automática destes dois recursos. Desta forma foi possível estender os rótulos de emoção para 2 729 conceitos. Foi também criado o maior recurso marcado com rótulos de emoção, bem como o primeiro léxico de emoção marcado quantitativamente.

[AM12] propôs um analisador de Sentimentos de Notícias baseado em léxico e orientado para comentários, denominado LCNSA (lexicon-based Comments-oriented News Sentiment Analyzer). Esta ferramenta pode lidar com o seguinte: a linguagem não padronizada utilizada actualmente por muitas pessoas, a deteção do alvo das opiniões dos utilizadores em cenários de múltiplos domínios e o desenho de um modelo de conhecimento linguístico com adaptabilidade de baixo custo. O sistema proposto é composto por um módulo de deteção automática de foco e um outro de AST, capazes de avaliar as opiniões dos utilizadores em tópicos de itens de notícias. Esses

módulos usam um léxico de taxonomia desenhado especificamente para análise de notícias. As experiências mostram que os resultados obtidos até agora são extremamente promissores.

[Nie14] examinou como a ANEW (Affetive Norms for English Word) e outros dicionários com palavras de sentimento funcionam para a deteção da força do sentimento em posts do microblog em comparação a um novo léxico (New ANEW), desenlvolvido especificamente para microblogs. Realizando experiências em postagens recolhidas na rede social Twitter, foi mostrado que a nova lista de palavras pode ter um desempenho melhor que a ANEW.

[SK14] descreveu um sistema de AST de última geração que deteta o sentimento de mensagens textuais informais curtas, como tweets e SMS. Esta ferramenta baseou-se em novos léxicos específicos para tweets e gerou um recurso lexical especial para as palavras negadas. O sistema foi classificado em primeiro lugar na tarefa compartilhada SemEval 2013 *Sentiment Analysis in Twitter.*

[CS14] estudou os fenómenos diacrónicos de dois diferentes domínios, a saber, sócio-político e desportivo, nos corpora da Google N-grams. A análise foi realizada com 761 e 34 palavras do domínio sócio-político e desportivo, respetivamente. A delimitação da época foi realizada na base na distribuição das palavras ao longo de certos períodos de tempo. Analisou-se também o fenómeno de mudança de opinião, usando a correlação entre as frequências de dois ou mais termos ao longo de um determinado período de tempo. Oito emoções foram encontradas com 14 000 palavras usando o WordNet-Affect (WNA) baseado em NRC Word-Emotion Association Lexicon (WNANRC) [SM10] e o Semeval Affective Text 2007 (SAT) [CS07]. A metodologia proposta pode ser estendida para prever mudanças futuras na sociedade como a correlação entre socialismo e capitalismo.

[TN15] desenvolveu o protótipo de um sistema que permite melhorar a precisão em AST. Esta ferramenta identifica corretamente as relações semânticas existentes entre as expressões de sentimento e o assunto abordado. No desenvolvimento deste protótipo foi feita a junção entre um analisador semântico, um analisador sintático e um léxico do sentimento. O resultado das experiências realizadas alcançou alta precisão (75–95%) na busca de sentimentos em páginas da Web e artigos de notícias.

[For15] propôs um léxico enriquecido especifico para a área de telecomunicações que poderá permitir às empresas deste domínio detetarem facilmente os clientes que pretendem abandonar os seus serviços. Este dicionário de palavras com sentimento, denominado DomainWords, contém 6 915 palavras. Os testes foram feitos em 800 documentos de texto compostos por opiniões e comentários de clientes de uma empresa de telecomunicações portuguesa. Esta ferramenta apresentou bom desempenho com uma taxa de acerto de 80.5%.

Uma abordagem baseada em classificação multi-rótulo (multilabel) para AST foi proposta em [SML15]. O protótipo proposto tem três componentes principais: segmentação de texto, extração de características e classificação multi-rótulo. As palavras segmentadas e as características de sentimentos foram baseadas nos seguintes dicionários de sentimento: Dalian University of Technology Sentiment Dictionary, National Taiwan University Sentiment Dictionary e HowNet Dictionary. Um estudo empírico detalhado sobre os três léxicos foi realizado para comparar os

seus desempenhos de classificação de sentimento. As comparações empíricas realizadas mostraram que o Dalian University of Technology Sentiment Dictionary tem o melhor desempenho entre os três diferentes dicionários de sentimento

[Has16] apresentou o SentiCircles, uma abordagem baseada em léxico para Análise de Sentimento em Texto no Twitter. Diferente das abordagens típicas baseadas em léxico, que oferecem polaridades fixas e estáticas do sentimento prévio das palavras, independentemente do seu contexto, o SentiCircles é capaz de atribuir uma orientação de sentimento específico ao contexto das palavras. Esta abordagem permite a deteção de sentimento tanto ao nível da entidade quanto ao nível do tweet, utilizando diferentes métodos. Avaliou-se a abordagem proposta em três conjuntos de dados do Twitter. Os resultados das avaliações realizadas mostraram que a abordagem proposta supera significativamente outros léxicos para deteção de sentimento relativamente às entidades e tweet.

Os léxicos genéricos, geralmente, não possuem termos curtos e informais pertencentes a um domínio e intervalo de tempo específico. Para completá-los com esse conteúdo, [dSG16] propôs um sistema para expansão de léxico que automaticamente extrai os termos mais atuais e relevantes em diferentes domínios e avalia o seu sentimento através do Twitter. A avaliação do sistema combinando com os métodos (ensemble) para a classificação de sentimento apresentou resultados que superam 19 métodos de AST.

[KdS17] apresentou o Unilex, um léxico orientado especificamente a tweets em Português Brasileiro. Nesta ferramenta, as palavras possuem um símbolo que a indica como negativa (-1), neutra (0) e positiva (1), de acordo com os sentimentos por elas transmitidos. Durante um estudo realizado, a partir de dados provenientes da Web, foi mostrado que este dicionário de palavras com sentimento apresenta melhor resultado em relação aos principais léxicos existentes.

A língua árabe está a expandir-se no mundo. Segundo a UNESCO, esta língua possui mais de 422 milhões de falantes nativos em cerca de 30 países, entre 1,6 biliões de muçulmanos em todo o mundo que a usam para realizar as suas orações diárias [IZ17]. A presença da língua árabe na internet cresceu com cerca de 6,091% nos últimos quinze anos [IZ17]. O número de documentos textuais em Árabe aumenta rapidamente. Isso exige a necessidade de melhorar as técnicas de processamento de texto nesta língua. Para suprir essa necessidade, [RAA14] apresentou um léxico de sentimento desenvolvido especificamente para avaliar texto de redes sociais em Árabe. Este dicionário de palavras com sentimento pode ser muito útil para minimizar as dificuldades ligadas à complexidade desta língua, à carência de publicações, e sobre tudo, à falta de ferramentas para AST.

### 2.2.2 Abordagens baseadas em Aprendizagem Automática

Até agora, existem poucas pesquisas realizadas sobre classificação de sentimento para documentos chineses. A fim de remediar essa deficiência, [ST08] apresentou um estudo empírico da AST aplicado em documentos chineses. Foram usados quatro métodos de seleção de características (MI: Mutual Information, IG: Information Gain, CHI: statistics e DF: Document Frequency) e cinco métodos de Aprendizagem Automática (Centroid Classifier, K-Nearest neighbor, Winnow Classifier, Naïve Bayes e SVM) num corpus chinês com tamanho de 1 021 documentos. Os resultados experimentais indicaram que o IG funciona melhor para seleção de termos sentimentais

e o SVM exibe melhor desempenho para classificação de sentimento. Além disso, descobriu-se que os classificadores de sentimento são severamente dependentes de domínios ou tópicos.

[CL09] propôs uma nova estrutura probabilística de modelagem baseada na Latent Dirichlet Allocation (LDA), chamada modelo de sentimento / tópico (JST), que deteta o sentimento e o tópico simultaneamente a partir de um texto. Ao contrário de outras abordagens de Aprendizagem Automática para classificação de sentimento que, muitas vezes, exigem corpora rotulados para treino de classificadores, o modelo proposto é totalmente não supervisionado. Portanto, fornece mais flexibilidades e pode ser mais facilmente adaptado a outras aplicações. Os resultados das experiências realizadas para avaliar o desempenho de JST, com base no conjunto de dados de revisão de filme, demonstraram que este modelo apresenta um bom desempenho na classificação de sentimento e os tópicos descobertos são, de fato, coerentes e informativos.

[ST09] atacou o problema de transferência de domínio, isto é, os problemas relacionados a redução do desempenho de classificadores de sentimento supervisionado quando estes são transferidos para novos contextos. Para selecionar recursos generalizáveis que ocorrem com frequência em ambos os domínios, foi proposta a medida efetiva FCE (Frequently Co-occurring Entropy). Para obter conhecimento dos dados de novos domínios, foi proposto o ANB (Adapted Naïve Bayes), uma versão de transferência ponderada do Naive Bayes Classifier. Os resultados experimentais indicaram que a abordagem proposta pode melhorar o desempenho do classificador de base como NTBC (Naid Bayes Transfer Classifier).

A língua árabe tem uma morfologia complexa. Por este motivo, até agora há indisponibilidade de ferramentas de análise morfológica para este idioma. [MKS10] apresentou e avaliou os algoritmos de stemming existentes em Árabe, como também implementou e acrescentou as ferramentas de análise morfológica árabe nas principais ferramentas de Aprendizagem Automática e Data Mining de código aberto, Weka e RapidMiner.

[SW11] realizou a classificação de subjetividade, considerando o melhor método de seleção de característica baseada em Fisher's discriminant ratio. As experiências foram realizadas em dois corpora chineses: revistas multidomínio e 11 revistas diferentes de marca de carros. Os conjuntos de recursos propostos juntamente com as palavras que aparecem em textos com sentimento positivo e negativo foram utilizados para o treino com Support Vector Machine (SVM). O resultado apresentou a Precisão de 86,6%.

A seleção de recursos é uma tarefa crítica na AST, pois, essas ferramentas se forem corretamente escolhidas, podem melhorar a classificação de opiniões em texto. [AC12] explorou a aplicabilidade de cinco métodos de seleção de recursos, geralmente empregues nas investigações em PLN (DF, IG, GR, CHI e Relief-F) e sete técnicas de classificação baseadas na Aprendizagem Automática (Naïve Bayes, Máquina de Vetores de Suporte, Máxima Entropia, Árvore de Decisão, K-Nearest Neighbor, Winnow, Adaboost) para análise de sentimento em conjunto de dados de análises de filmes online. Esta investigação demonstrou que a seleção de recursos melhora o desempenho da classificação baseada em sentimento, mas depende do método adotado e do número de recursos selecionados. Os resultados experimentais mostraram que o Gain Ratio oferece o melhor desempenho para a seleção de recursos sentimentais.

[VN13] baseou-se na busca de diferentes métodos para melhorar a Precisão do classificador Naive Bayes em AST. Ao longo desse estudo, treinando e testando um modelo simples de Naive Bayes, desenvolveu-se um classificador de sentimento altamente preciso e rápido. Este algoritmo resultou da combinação de métodos como o tratamento efetivo da negação, palavra n-gram e a seleção de características por informações mútuas. As experiências realizadas sobre o popular conjunto de dados de filmes da IMDB obtiveram mais 85% de Precisão.

[AO13] apresentou um novo método de AST aplicada ao Facebook, que a partir de mensagens escritas por utilizadores, suporta:

a) extrair informações sobre a polaridade de sentimento dos utilizadores (positivo, neutro ou negativo);

b) modelar essa polaridade e detetar mudanças emocionais significativas.

Este método de classificação foi implementado no SentBuk, um aplicativo que recupera mensagens escritas por utilizadores do Facebook e as classifica de acordo com a sua polaridade. Os resultados obtidos dessa experiência mostraram que o método é viável para AST em Facebook com alta precisão (83,27%).

[GW14] realizou uma avaliação comparativa do desempenho de três métodos de aprendizagem conjunta populares (Bagging, Boosting e Random Subspace). Este estudo baseou-se em cinco máquinas de base (Naive Bayes, Maximum Entropy, Decision Tree, K Nearest Neighbor e Support Vector Machine) para classificação de sentimento. Além disso, dez conjuntos de dados de análise da opinião pública foram investigados para verificar a eficácia da aprendizagem conjunta para AST. Com base nas experiências realizadas, os resultados revelam que os métodos de aprendizagem conjunta melhoram substancialmente o desempenho de máquina de base para a classificação do sentimento. Dentre os três métodos ensemble, o Random Subspace apresenta melhores resultados.

Os métodos baseados em Aprendizagem Automática para classificação de sentimento apresentam excelente desempenho. Apesar disso, a maioria das pesquisas existentes neste domínio está centrada na extração de características lexicais e características sintáticas, enquanto as relações semânticas entre as palavras são ignoradas. Em [DZ15], a fim de obter as características semânticas, foi proposto um método para classificação de sentimento baseado em word2vec e SVMperf. Realizaram-se as experiências no conjunto de dados de comentários chineses sobre produtos de vestuário. Os resultados experimentais apresentaram taxas de acerto acima de 80%.

[PC15] descreveu um sistema de AST baseado em Twitter, desenvolvido combinando classificadores baseados em regras e Aprendizagem Supervisionada. Os resultados de treinos realizados em Support Vector Machine (SVM) mostram que as regras criadas podem ajudar a refinar as previsões do SVM.

[Kau16] propôs dois métodos direcionados a duas abordagens fundamentais:

a) análise de sentimento baseada em aspetos;

b) atribuição de polaridade.

Para a análise de sentimento baseada em aspetos, o método foi desenvolvido a partir de algoritmos de classificação e permite identificar expressões que mencionem os aspetos das entidades que se encontra num texto. Para a atribuição de polaridade, desenvolveu-se, a partir de modelos de Aprendizagem Automática, um método independente de recursos linguísticos que utiliza 24 atributos e que pode ser aplicado sobre dados com ruído.

[LD16] teve como foco a aplicação das técnicas de Aprendizagem Automática na descoberta e análise de conteúdos sentimentais de críticas de filmes e avaliações de hotéis. O artigo discutiu elaboradamente dois algoritmos de Aprendizagem Supervisionada: K-Nearest Neighbor (K-NN) e Naive Bayes; e comparou os valores das suas Precisões aos seus Recall. Verificou-se que, no caso de críticas de filmes, Naive Bayes apresenta resultados melhores do que o K-NN, mas, para avaliações de hotéis, os algoritmos apresentam menores precisões, quase as mesmas.

[KC09] incorporou num modelo de previsão de rotatividade para um negócio de assinatura de jornais, as emoções expressas nos e-mails *clientes / empresas* . Para as experiências, 18 331 e-mails de um jornal belga foram comprados. O Linguistic Inquiry and Word Count (LIWC [JP07]) foi utilizado para criar uma lista de 690 palavras positivas e 1 347 palavras negativas, o que ajudou na classificação do conteúdo dos e-mails. As ferramentas: Logistic Regression (LR), Support Vector Machine (SVM) e RF foram utilizadas para classificação de sentimento. Os resultados mostraram que RF supera os outros dois classificadores.

### 2.2.3 Outras abordagens

O número de comentários deixados por clientes que fazem compra online cresce rapidamente. Para simplificar a sua leitura na parte dos clientes e também ajudar os fabricantes a acompanhar as opiniões dos clientes sobre os seus produtos, [HL04] propôs um conjunto de técnicas que permitem caracterizar e identificar o produto sobre o qual os clientes expressaram as suas opiniões, as frases com polaridade positiva e negativa assim como o resumo da informação descoberta.

A negação é uma construção linguística muito comum que afeta a polaridade, [MWDK10] apresentou uma pesquisa sobre o papel da negação na Análise de Sentimentos em Texto com várias abordagens. O foco principal do trabalho foi a deteção, numa frase, do termo da negação e o do seu âmbito.

A análise automática de opiniões on-line envolve um profundo entendimento do texto em linguagem natural por máquinas, do qual ainda estamos muito distantes. Para este fim, [Cam13] apresentou uma nova abordagem em Processamento de Linguagem Natural que é a análise de sentimento ao nível do conceito. Perante a enorme quantidade de texto não estruturado publicado na web. Esta abordagem poderá ser muito útil para passagens mais eficientes de dados não estruturados para estruturados, processáveis por máquina, potencialmente em qualquer domínio. Assim as limitações encontradas atualmente nas abordagens semânticas poderão ser superadas.

A análise de sentimento dependente do tópico de um documento ainda é um território relativamente desconhecido. O método utilizado atualmente para realizar esta tarefa, geralmente,

não apresenta resultados satisfatórios. Para incentivar o desenvolvimento de métodos para medir sentimentos específicos de tópicos em documentos, [TMSA14] descreveu o processo de anotação e avaliação da qualidade de conjuntos de dados a partir de um corpus formado de 297 documentos e mais de 9 000 frases, utilizando várias medidas como Kappa statistic, Intraclass Correlation Consistency, Intraclass Correlation Agreement, Average Percentage Agreement, Finn-Coefficient.

O rápido crescimento dos dados gerados pelas aplicações web e de média social faz com que surjam novos paradigmas na geração de conhecimento. [AMR14] apresentou o CESA (Crowd Explicit Sentiment Analysis), uma nova abordagem para a classificação de polaridades em texto proveniente de redes sociais. Semelhante à análise semântica explícita, as publicações do microblog são indexadas por uma coleção predefinida de documentos. No CESA, esses documentos são construídos a partir de expressões emocionais comuns em fluxos sociais. Desta forma, o texto é projetado para sentimento ou emoção. O seu design simples permite a construção de classificadores de polaridade em diferentes idiomas e domínios. O sistema foi avaliado com conjuntos de dados em Inglês e Espanhol.

[Oli15] serviu para verificar se a Análise de Sentimento em Texto pode trazer contribuições para as práticas da gestão social colaborando com a consolidação de um modelo de gestão público mais democrático. Para realização das experiências, utilizou-se, como data set, os tweets baseados em opinião dos cidadãos sobre os principais programas do governo federal do Brasil nos últimos anos. Essa investigação revelou que, utilizando as técnicas de AST para determinar o nível de satisfação da sociedade civil sobre diferentes temas de interesse público, é possível melhorar a gestão de projetos sociais e construir um estado mais democrático.

[Bel17] baseou-se na relação existente entre a Educação à Distância, Linguística, e Processamento de Linguagem Natural para criar regras que permitem detetar a negação em texto e resolver os problemas que esta tem causado para AST. Utilizando um corpus construído em contexto de ensino à distância com base nos relatos diários e fóruns dos alunos, criou-se 11 regras linguístico-computacionais que poderão contribuir para que um sistema computacional possa localizar os fenómenos da negação em textos e verificar a existência de inversões de polaridade e emoção.

### 2.2.4 Análise do estado da arte

Em estado da arte, foi feito um mapeamento das pesquisas já realizadas no nosso campo de estudo. Como foi visto, o avanço que temos em Análise de Sentimento em Texto é resultado de contributos de vários investigadores. Estes desenvolveram novas ferramentas e outros aperfeiçoaram as técnicas já existentes. Apesar desse progresso, é possível notar que poucos trabalhos têm aplicação em línguas diferentes de Inglês, pior ainda em domínios específicos. Os gráficos abaixo resume estatisticamente as publicações encontradas.


Figura 2.2: Investigação por área de aplicação.

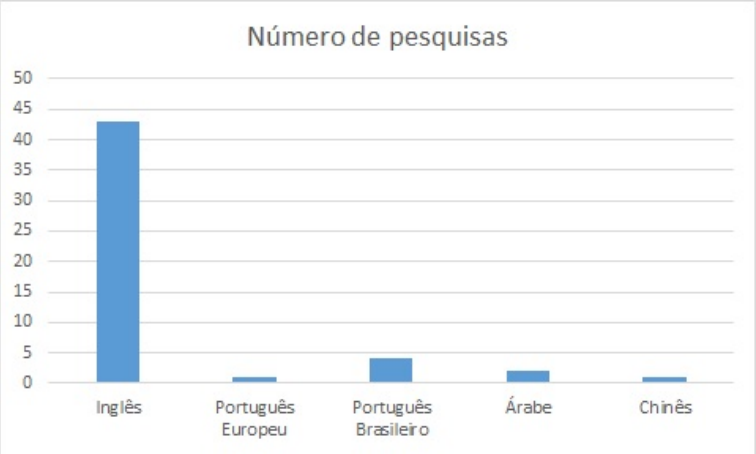


Figura 2.3: Número de pesquisa por idioma.

Quais 90% dos trabalhos encontrados são genéricos. Da mesma forma, o Inglês é a língua visada por cerca de 80% dos trabalhos.

Deste modo, nota-se que ainda há muito espaço a explorar neste campo de PLN. Falta investigar bastante em contexto específico. As futuras investigações devem abranger também outras línguas, para que as organizações do mundo inteiro, independentemente dos seus idiomas de trabalho, desfrutem das vantagens da análise automática de sentimento em texto.

## 2.3 Sumário

Este capítulo relatou os aspetos teóricos dessa dissertação: conceituou a AST e áreas relacionadas, apresentou os aspetos linguísticos da investigação e na parte final debruçou sobre os trabalhos relacionados. O próximo capítulo trata dos aspetos representacionais e computacionais.

# Capítulo 3

# Caraterização do sistema

Após ter apresentado as propostas de solução para a situação em análise nesta dissertação, vem a fase de traduzi-las em código tratável pelo computador. Para tal, surge o Capítulo 3, que, para além deste objetivo, trata de apresentar os principais aspetos do sistema pretendido. Assim sendo, este capítulo é composto de 3 secções: análise linguístico-computacional, análise computacional e sumário.

## 3.1 Análise linguístico-computacional

Tendo como referência as propostas apresentadas no capítulo anterior, esta secção trata de apresentar e traduzir em linguagem algorítmica, as soluções adotadas nesta dissertação para realizar o processo de Análise de Sentimento em Texto (AST).

### 3.1.1 Algoritmo

De acordo com o popular livro didático de algoritmos [TH02], um algoritmo é qualquer procedimento computacional bem definido que leva algum valor, ou conjunto de valores, como entrada e produz algum valor, ou conjunto de valores como saída. Em outras palavras, os algoritmos são como guias para realizar uma determinada tarefa bem definida. Em construção civil, o algoritmo pode ser comparado ao projeto elaborado por arquiteto, cuja implementação está a cargo dos pedreiros. Imagina-se a construção civil sem a existência dos arquitetos, claro que os pedreiros não seriam capazes de edificar as maravilhosas obras arquitetónicas que observamos atualmente. Nesta dissertação, utilizámos os algoritmos para partilhar as soluções encontradas para os diferentes problemas em análise e permitir que elas sejam entendidas por qualquer programador, independentemente da sua linguagem de programação. Esses algoritmos são representados em pseudocódigo, pelo fato de ser uma representação muito próxima à linguagem humana e fácil de ser traduzida para uma linguagem de programação.

### 3.1.2 Modularização e Programação Orientada a Objeto

Quanto mais complexa é uma tarefa em algoritmia, mais difícil será a sua compreensão, assim como a sua manutenção. Para evitar esses problemas, os programadores recorrem à modularização. Este é um processo que consiste em dividir a tarefa a realizar em pequenas tarefas (módulos), nas quais cada uma delas tem a responsabilidade de realizar uma etapa específica da tarefa principal. A modularização é a técnica que utilizámos para resolução da maior parte dos problemas nesta dissertação.

Para aproximar o mundo digital ao mundo real, associámos à modularização a Programação Orientada a Objeto (POO). Este paradigma de programação enfatiza a criação de classes que encapsulam os dados e algoritmos a manipular. A grande vantagem da POO é a possibilidade de

reutilizar os códigos de outras classes previamente programadas (Herança). A utilização desses dois recursos (modularização e POO) permite também, poupar o esforço e ganhar o tempo no processo de desenvolvimento de software.

### 3.1.3 Proposta de solução

Para o desenvolvimento do sistema proposto, propusemos uma abordagem que consiste em criar um projeto com três classes principais: *Palavra*, *Sentenca* e *Texto*. Os seus nomes indicam o que elas representam. A classe *Sentenca* é representada através de uma lista ligada de elementos da classe *Palavra*. Da mesma forma, a classe *Texto* é representada através de uma lista ligada de elementos do tipo *Sentenca* . Essas classes são compostas por vários métodos, nas quais são implementadas as diferentes tarefas do processo de AST mencionadas neste trabalho e não só. Sendo a frase a unidade fundamental de AST, as próximas subsecções tratam de descrever os algoritmos dos principais métodos desta classe.

### 3.1.4 Pré-processamento

Algorithm 1 - pré-processamento.

1: Início
2: Segmentar a sentença
3: Corrigir erros ortográficos
4: Transformar letras maiúsculas em minúsculas
5: Eliminar Stop words
6: Fim

O pré-processamento obedeceu à configuração proposta na Subsecção 2.1.7. Isto é, teve as seguintes fases: tokenização e normalização (correção de erros ortográficos, substituição de letras maiúsculas por minúsculas e supressão de stop words). Para facilitar a realização das Tarefas 3 e 5 (do Algoritmo 1), criou-se um ficheiro de texto, com uma lista de cerca de 400 abreviaturas junto com as suas formas normalizadas e outro com as principais stop words. A Tabela 3.1 apresenta um estrato desses ficheiros.

| Stop words | ao | Abreviaturas | V.Exa. = Vossa Excelência |
|---|---|---|---|
| | sobre | | zool.= zoologia |
| | sua | | zootec. = zootecnia |
| | depois | | hist.= história |
| | ainda | | i.e. = isto é |
| | para | | Cia. = companhia |
| | pela | | cód. = código |
| | com | | abrev. = abreviatura |

Tabela 3.1: Stop words e ortográfia.

### 3.1.5 Etiquetagem morfológica e Atribuição da polaridade ao nível da palavra

Neste projeto, a classe *Palavra* tem dois atributos principais: *polaridade* e *claGram*, que representam, respetivamente, a polaridade e a classe gramatical da palavra. Deste modo, o part-of-speech tagging é feito via programação através de uma ferramenta cujos detalhes encontram-se na Subsecção 3.2.2
O Algoritmo 2 representa a função responsável para a polarização de vocábulos.

Algorithm 2 - PolaridadePalavra

```
1: Início
2: for (w ∈ Frase) do
3:    if lema(w) ∈ LéxicoSentimento then
4:       w.Polaridade←lerPolaridade(lema(w),LéxicoSentimento)
5:    else
6:       w.polaridade←0
7: Fim
```

No Passo *2* deste algoritmo, percorrem-se todas as palavras da frase. No Passo *3*, é feita a lematização das palavras e é verificada se estas constam no léxico. Caso conste, a polaridade da palavra é a sua correspondente no léxico (Passo *4*), caso contrário, a palavra recebe o valor 0 como polaridade (Passo *6*). A lematização é feita através de uma função específica, herdada da biblioteca *Mxpost* (ver Subsecção 3.2.2)

### 3.1.6 Previsão da polaridade em frase.

A frase é uma unidade linguística com sentido completo, que contém pelo menos um verbo principal, sendo delimitada na escrita por letra maiúscula, no início, e no fim, por um sinal de pontuação [Pe18]. Neste trabalho, a abordagem escolhida para previsão de sentimento em frase é a proposta em [AE06], descrita na Subsecção 2.1.9. Deste modo, a polaridade da frase, foi determinada pela soma de polaridades das suas palavras.

Algorithm 3 - PolaridadeFrase

```
1: Início
2: polF ← 0
3: preprocessamento
4: polaridadePalavra
5: for (w ∈ Frase) do
6:    polF← polF+w.polaridade
7: retorna polF
8: Fim
```

O Algoritmo 3 representa uma função que retorna o valor da polaridade de uma frase. No Passo 2, declarou-se uma variável para acumula o valor da polaridade. Nos Passos 3 e 4, chamaram-se, respetivamente, as funções de pré-processamento da frase e da polarização de palavras. No Passo 6, soma-se as polaridades de todas as palavras da frase.

### 3.1.7 Tratamento de expressões idiomáticas

Na Subsecção 2.1.10, mostrámos como a presença das expressões idiomáticas nas frases influenciam o processo de Análise de Sentimento em Texto. Na mesma parte deste trabalho, demonstrámos como esses conjuntos de palavras podem reduzir significativamente a probabilidade da correta classificação de sentimentos expressos nas frases que as contêm, se não forem

bem tratados. No entanto, até agora, o algoritmo implementado para previsão de sentimento em frase não trata desta questão. Logo, há necessidade de adapta-lo a esta realidade. Para tal, implementámos as seguintes regras:

**Regra 1: No processo da previsão de sentimento ao nível da frase, o cáculo da polaridade deve ser precedido pela verificação da existência de expressões idiomáticas.**

**Regra 2: Em AST, à expressão idiomática é atribuida uma polaridade única, os seus elementos constituintes são ignorados.**

O Algoritmo 4 representa uma função que verifica se uma frase é expressão idiomática.

Algorithm 4 - IsExprIdiom

```
Início
for (expIdio ∈ listaExpIdiom) do
   if (frase=expIdio) then
      retorna true
   retorna false
Fim
```

### 3.1.8 Tratamento de amplificadores e atenuadores

O tratamento de adverbios de intensidade, neste projeto, baseou-se na solução abordada em [PL06] (ver Subsecção 2.1.12), na qual propõe-se a alteração da intensidade da polaridade do termo que o amplicador/atenuador modifica. As regras implementadas para a aplicação desta abordagem foram as seguintes:

**Regra nº1: Numa frase, a polaridade da próxima palavra com sentimento após o amplificador é multiplicada por 1.6.**

**Regra nº2: Numa frase, a polaridade da próxima palavra com sentimento após o atenuador é multiplicada por 0.6.**

Como ilustração, comprovamos as frases seguintes:

a) *Joana é muito querida.*

b) *O comportamento menos extravagante do elenco foi notável.*

A polaridade de *querida*, no Sentilex-Pt, é 1. Entetanto, no processo de classificação da frase onde este adjetivo se encontra, lhe séra atribuido o valor 1*1.6=1.6, por ser a primeira palavra com sentimento depois do amplificador *muito*. Quanto a *extravagante*, a sua polaridade passará de -1 para -0.6, por ser antecedido pelo atenuador *menos*.

Neste trabalho, os amplificadores e atenuadores, estão guardados em dois ficheiros de texto distintos que, na prática, serão carregados no sistema e utilizados no processamento.

O Algoritmo 5 ilustra as diferentes etapas da função *AmpliAtenuador* desenvolvida, especificamente para tratar dos amplificadores e atenuadores.

Algorithm 5 - AmpliAtenuador

```
1: Início
2: for ( w ∈ frase) do
3:    if (w ∈ listAmplif) then
4:       w←w.next()
5:       while (w.polaridade=0) do
6:          w←w.next()
7:       w.polaridade←w.polaridade*1.6
8:    else
9:       if (w ∈ listAten) then
10:         w←w.next()
11:         while (w.polaridade=0) do
12:            w←w.next()
13:         w.polaridade←w.polaridade*0.6
14: Fim
```

### 3.1.9 Tratamento da negação

A solução adotada para a modelação da negação baseou-se nas diferentes propostas apresentadas na Subsecção 2.1.11, nas quais é invertido o sinal da polaridade do termo modificado pela presença deste fenómeno linguístico. Deste modo, implementaram-se as seguintes regras:

**Regra 1: Numa frase, a polaridade da primeira palavra com sentimento após o termo de negação é multiplicada por -1.**

O exemplo a baixo ilustra a aplicação desta regra:

*Porque nós sabemos que a nossa herança de diversidade é uma força, não fraqueza.*

Neste caso, a polaridade de *fraqueza* passará de -1 para 1. No entanto, esta regra não é universal; ao longo desta investigação, deparámo-nos com situações como a do exemplo seguinte:

*O preço do baril do petróleo, este ano, não foi muito elevado.*

reescrevendo, a mesma frase sem o termo de negação *não*, teremos:

*O preço do baril do petróleo, este ano, foi muito elevado*

Nota-se que, com ou sem *não*, nesta frase, o sentimento é sempre positivo. A diferença é que quando entra o adverbio de negação, o grau de positividade é reduzido. Logo conclui-se que o adverbio de negação *não*, nesta situação, está simplesmente a anular a função do amplificador *muito*. Perante esta constatação, propusemos uma segunda regra para lidar com a negação.

**Regra 2: Quando a próxima palavra com sentimento, depois do termo de negação, é precedida de um amplificador/atenuador, não se aplica a Rega 1, mas simplesmente anula-se a função deste adverbio de intensidade.**

Neste trabalho, as duas regras foram implementadas numa função específica, representada no Algoritmo 6.

Na semelhança de advérbios de intensidade, criou-se um ficheiro de texto onde estão armazenados os advérbios de negação.

Algorithm 6 - Negacao

```
1: Início
2: for (w ∈ frase) do
3:    if w ∈ listTermoNeg then
4:       w←w.next()
5:       while (w.polaridade=0) do
6:          w←w.next()
7:       if (w.previous()∉ listAmp)⋀(w.previous()∉ listAten) then
8:          w.polaridade←w.polaridade*-1
9: Fim
```

### 3.1.10 Função para previsão de sentimento em frase

O Algoritmo 7 representa a função atualizada para o cálculo da polaridade em frase. Este método é o núcleo da classe *Sentenca*, pois, nele são chamados todos os outros, implementados nesta secção e não só.

Algorithm 7 - PolaridadeFrase

```
1: início
2: polF←0
3: Pré-processamento
4: PolaridadePalavra
5: AmpliAtenuador
6: Negacao
7: if  (IsExpIdiom) then
8:    retorna ExpIdiom.polaridade
9: for ExpIdiom ∈ ListaExpIdiom do
10:    if  ( ExpIdiom⊆ Frase) then
11:       PolF←ExpIdiom.polaridade
12:       Frase←Frase − ExpIdiom
13: for (w ∈ Frase) do
14:    polF← polF+w.polaridade
15: retorna polF
16: Fim
```

No Passo 7 do pseudocódigo acima, verifica-se se a frase é uma expressão idiomática através da função IsExpIdom; caso seja, o Passo 8 termina a função retornando a polaridade desta expressão idiomática. No Passo 10, verificar-se se parte da frase é expressão idiomática; caso essa condição seja verdadeira, a polaridade da frase é determinada pela soma da polaridade da expressão idiomática e das restantes palavras (Passos 11,12,13,14).

## 3.2 Análise computacional

Após ter criados os algoritmos necessários para resolver o problema em estudo, o passo seguinte constituiu no desenvolvimento do sistema informático. Esta secção dedica-se à descrição dos principais aspetos deste software.

### 3.2.1 Modelo de processo do desenvolvimento do sistema

O processo de desenvolvimento de software é um conjunto de atividades, parcialmente ordenadas, com a finalidade de obter um produto de software [wik17]. Este mecanismo é importante

para se obter um software de qualidade.

Os processos de desenvolvimento de software são agrupados em vários modelos. Entre eles: o modelo em cascata, em protótipo, em espiral, de entrega incremental, o Rational Unified Process (RUP), etc. Neste projeto, de acordo com o paradigma de programação proposto, o modelo escolhido é o baseado em componentes. Este modelo de desenvolvimento de software consiste em reutilizar os módulos de sistemas existentes, em novos projetos do mesmo contexto, isto é, baseado no paradigma de Programação Orientada a Objeto. As vantagens deste modelo estão na redução do prazo de ciclo de desenvolvimento e de custo do projeto, como também no aumento da produtividade.

### 3.2.2 Ferramentas

A codificação do sistema exige preparar o computador com um conjunto de ferramentas de software. Para tal é necessário estudar e escolher as melhores tecnologias que permitem obter uma aplicação de melhor qualidade. Nos próximos passos descrevemos as ferramentas utilizadas neste projeto, assim como as razões pelas quais foram escolhidas.

**Java**: Esta linguagem de programação foi projetado para atender aos desafios do desenvolvimento de aplicativos no contexto de ambientes distribuídos heterogéneos em toda a rede [JG96]. Um dos seus desafios é a entrega segura de aplicativos que consomem o mínimo de recursos do sistema e que podem ser executados em qualquer plataforma de hardware e software [JG96]. As características do Java incluem uma linguagem simples que permite o desenvolvimento de sistemas informáticos seguros, de alto desempenho, altamente robustos e confiável.

Java é adequado para o paradigma de programação proposto neste trabalho, pois, esta linguagem é projetada para ser orientada a objetos. Ela fornece uma plataforma de desenvolvimento limpa e eficiente baseada em objetos. A linguagem Java disponibiliza bibliotecas com várias funcionalidades: os tipos de dados básicos, a manipulação de ficheiros (I/O), etc.

**Java Development Kit**: Esta ferramenta é um ambiente de desenvolvimento de software usado para desenvolver aplicativos na linguagem de programação Java. Este recurso inclui, um interpretador (java), um compilador (javac), uma biblioteca (jar), um gerador de documentação (javadoc), etc. Em suma JDK possui todas as ferramentas necessárias para compilar, depurar e executar aplicativos escritos na linguagem Java.

**HultigLib**: Este recurso é uma biblioteca desenvolvida pelo *Centro de Tecnologia da Linguagem Humana e Bioinformática*. Este grupo de investigação do Departamento de Informática da Universidade da Beira Inteiror tem como objetivo, desenvolver investigação fundamental e aplicada nas áreas da Tecnologia da Linguagem Humana e Bioinformática [hul18].

Esta biblioteca reúne um conjunto de ferramentas de processamento de texto, escritas em linguagem Java. O pacote *sumo* da HultigLib possui classes e métodos com estruturas de dados semelhantes aos propostos nesta dissertação. As suas principais classes são: *Word, Sentence* e *Text*. A medição da distância entre palavras ou frases, a segmentação de texto em frases, são algumas das várias tarefas realizadas pelas funções dessas classes. Deste modo, desfrutando das

vantagens da herança em POO, reutilizámos os códigos das classes desta biblioteca, necessários para este projeto.

**Mxpost**: Recorremos a esta ferramenta neste projeto especificamente para a etiquetagem morfológica e para a lematização. Nesta biblioteca encontra-se uma classe com um método que retorna a forma canônica de palavras, como também determina as suas classes gramaticais. O método apresenta algumas limitações quando se trabalha com palavras polissémicas.

O desafio aqui, é de desenvolver um método que acrescente a esta tarefa, a análise sintático-semântica para superar esta limitação, retornando a classe gramatical da palavra mediante a função que este desempenha na frase em que se encontra. Os exemplos abaixo ilustram a situação de palavras polissémicas:

a) *João e Lourenço foram ao **banco** levantar dinheiro.*

b) *O **banco** no qual Henriqueta sentou, está no centro do jardim público.*

c) *Eu **canto** todas as manhãs no **canto** da casa.*

A palavra *banco* dos exemplos da alinha *a)* e *b)*, apesar de ter significados diferentes, a sua classe gramatical (substantivo) não mudou. Nem sempre acontece assim, no exemplo da alinea *c)*, a palavra *canto*, numa mesma frase aparece duas vezes com classes gramaticais diferentes. O primeiro *canto* é um verbo, portanto, o segundo é um substantivo.

**NetBeans**: O fato deste ambiente integrado para desenvolvimento em Java ser gratuito, de código aberto e ter uma grande comunidade de utilizadores e programadores em todo o mundo, foi a principal razão da sua escolha neste projeto.

**Gitub**: Esta ferramenta permite o armazenamento na cloud de repositórios para desenvolvimento de softwares. Ela pode ser considerada como uma rede social de programadores. Este recurso permite criar repositórios, públicos como privados, propícios para manter os projetos guardados na cloud, evitando assim, o risco de perde-los caso aconteça algo na máquina física. Além disso, com o Github é possível colaborar com a maioria de projetos open source existentes atualmente na web. Neste trabalho, a utilização deste recurso permite também que o nosso projeto esteja disponível para a comunidade.

### 3.2.3 Diagrama de classes

O diagrama de classes é muito importante quando se trabalha num sistema orientado a objetos, pois, este artefacto da UML oferece uma representação visual da modelagem de classes e as relações que existem entre elas. Assim sendo a Figura 3.3 apresenta as principais classes do projeto, assim como os seus principais atributos e métodos.

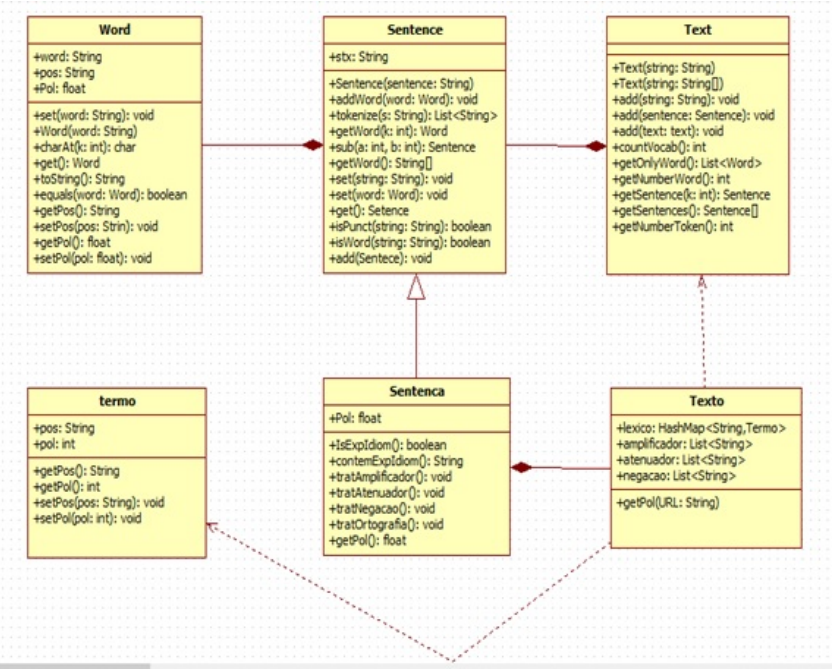


Figura 3.1: Diagrama de classes.

### 3.2.4 Requisitos funcionais

| Requisito | Descrição |
|---|---|
| Classificar frases estruturadas. | classifica um conjunto de frases bem estruturadas. |
| Classificar frases não estruturadas. | Organiza o texto do ficheiro indicado pelo utilizador em frases; e atribuir polaridades a cada uma delas. |
| Classificar uma única frase. | Recebe como entrada uma frase e determina a sua polaridade. |
| Classificar uma única palavra. | Classifica o vocábulo inserido pelo utilizador quanto ao seu sentimento. |
| Obter lema de uma palavra. | Recebe como entrada um verbo conjugado e retorna a sua forma canónica. |
| Realizar a etiquetagem morfológica de uma palavra. | Dertermina a classe gramatical de um vocábulo assim como o seu género. |

Tabela 3.2: Requisítos funcionais do sistema.

### 3.2.5 Descrição do sistema

Neste trabalho, desenvolveu-se um sistema para Análise de Sentimento em Texto, chamado Sentisoft. O desenvolvimento desta ferramenta baseou-se nos pormenores da proposta de solução apresentada neste trabalho: modelo do processo de desenvolvimento, paradigma de programação, algoritmos, linguagem de programação, biblioteca, etc.

### 3.2.6 Formulário principal

A Figura 3.4 representa a primeira janela de interação com o utilizador, a partir da qual é possível aceder ao menu principal.

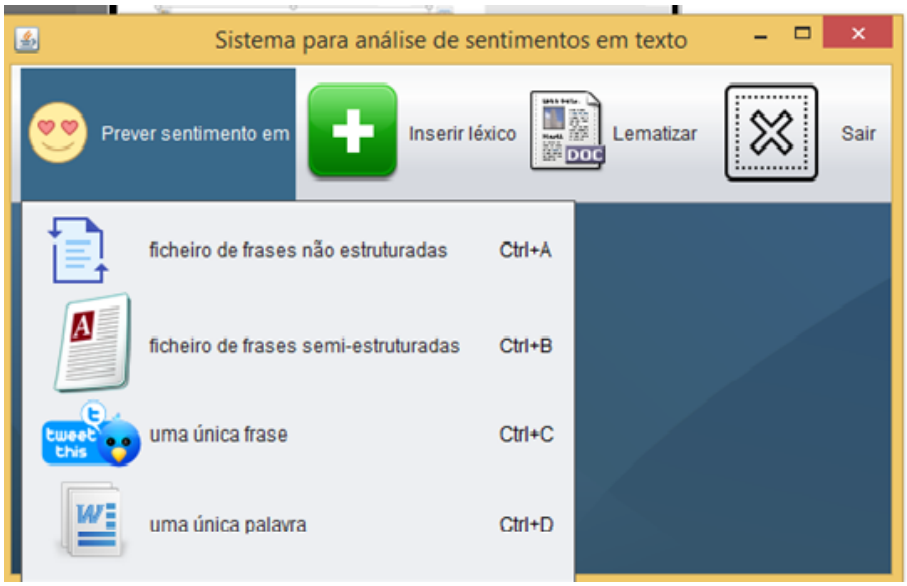


Figura 3.2: Formulário principal.

### 3.2.7 Formulário do resultado de classificação de frases

A Figura 3.5 representa um formulário que visualiza o resultado da polarização de uma lista de frases de um determinado corpus, escolhida pelo utilizador.

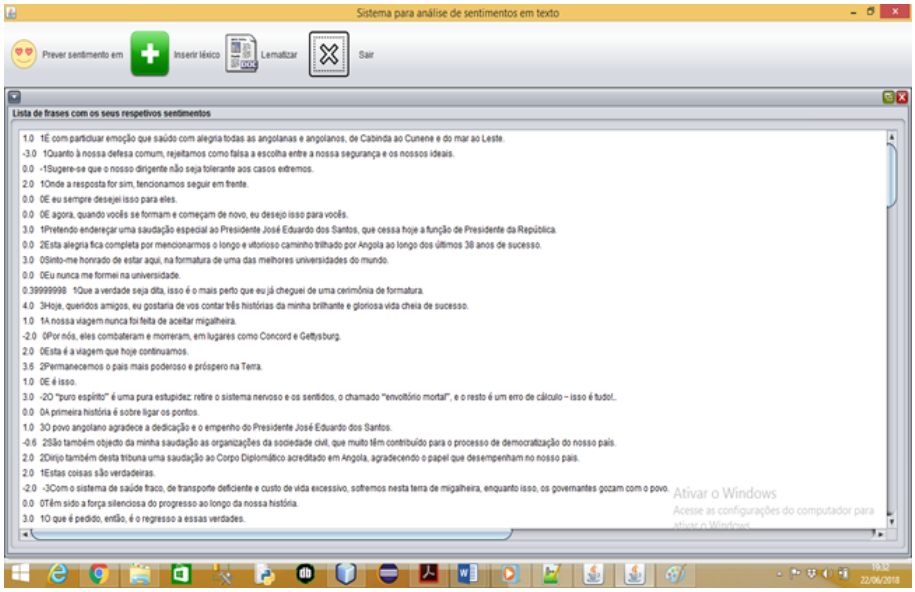


Figura 3.3: Formulário da classificação de frases.

## 3.3 Sumário

O presente capítulo fez uma descrição detalhada da aplicação desenvolvida neste trabalho para análise de sentimento em texto. No Capítulo 4, esta ferramenta serviu de base para avaliar o desempenho dos léxicos geénricos em Português, quanto à AST na área da Economia.

# Capítulo 4

# Experiências e Resultados

Na sequência da dissertação, e em resultado do trabalho desenvolvido, foram realizadas diversas experiências, sendo aqui analisados e discutidos os resultados obtidos. Deste modo, o presente capítulo está repartido em quatro secções: experiências com texto genérico, experiências com texto do domínio económico, considerações e sumário.

## 4.1 Experiências com texto genérico

As experiências com texto genérico foram realizadas com o objetivo de avaliar a capacidade do sistema desenvolvido na previsão de sentimento em texto. Esta avaliação é de grande importância para garantir a qualidade desta ferramenta informática. Isto é, assegurar que o sistema desempenha satisfatoriamente as tarefas para as quais foi desenvolvido. Durante esta operação, foram identificadas as falhas no seu funcionamento e proporcionadas as respetivas soluções.

Existem várias formas de testar o desempenho de softwares. Neste capítulo foram utilizadas duas delas: o teste de caixa branca e o de caixa preta. O teste de caixa branca foi realizado ao nível do código fonte, verificando os fluxos de execução das estruturas de decisões, dos ciclos, e outros. Com o teste de caixa preta executámos o sistema, várias vezes, com diferentes entradas e observámos atentamente os resultados. O foco desta operação foi os resultados de frases simples, as que contêm expressões idiomáticas, advérbios de intensidade e frases negativas.

De acordo com [VF11], avaliar consiste em comparar o real com o esperado, o que se planeou com o que se concretizou. Deste modo, o teste do sistema desenvolvido foi feito utilizando como entradas, frases antecipadamente classificadas; e para cada uma delas, foi feita a comparação do resultado do sistema ao resultado real. Quanto mais próximos estes fossem, melhor era a avaliação. A realização desta tarefa necessitou de um dataset de frases previamente polarizadas. A anotação manual de dados para a criação de um dataset é uma atividade com alto custo de tempo, por isso, vários investigadores preferem utilizar recursos pré-existentes [Sil15].

Para o caso concreto desta investigação, o fato da Análise de Sentimento em Texto ser uma área de estudo muito nova torna difícil encontrar um conjunto de frases em Português com sentimentos previamente atribuídos. Diante disso, foi necessário elaborarmos o nosso próprio conjunto de dados de teste. Assim, extraímos e polarizámos diferentes frases em corpora como: o discurso da tomada de posse do presidente angolano, o discurso de Steves Job na universidade de Stanford, músicas de Michael Jackson, mensagem do fundador do grupo chinês Alibaba, o protesto contra o aumento de preço de combustível em São Paulo, o fim de acordo fiscal entre a Finlândia e Portugal, etc. Por conseguinte, criaram-se dois datasets, um composto de 151 frases genéricas e outro de 64 frases da área económica, cujas listagens se encontram no Anexo A.2 e A.3, respetivamente. Já que a base de dados lexical disponível para o teste do sistema é

composta por léxicos genéricos (Sentilex-Pt e OpLexicon), nesta secção trabalhou-se somente com o dataset de frases genéricas.

### 4.1.1 Experiências com frases genéricas simples

Neste trabalho, frase simples refere-se àquela que não é ambígua, não é negativa, não tem advérbios de intensidade e não contem expressões idiomáticas. A Tabela 4.1 ilustra uma amostra de 10 frases desta categoria, escolhidas no conjunto das várias utilizadas.

| Nº | Frases | Pol. real | Pol. Sistema Sentilex-Pt | Pol. Sistema OpLexicon |
|---|---|---|---|---|
| 1 | Esta alegria fica completa por mencionarmos o longo e vitorioso caminho trilhado por Angola ao longo dos últimos 38 anos de sucesso. | 2.0 | 3.0 | 2.0 |
| 2 | Os médicos me disseram que aquilo era certamente um tipo de câncer incurável. | -1.0 | -1.0 | -1.0 |
| 3 | O meu médico aconselhou-me a ir para casa e arrumar as minhas coisas | -1.0 | 0.0 | 0.0 |
| 4 | Dirijo também nesta tribuna uma saudação ao corpo diplomático acreditado em Angola, agradecendo o papel que desempenha no nosso país. | 1.0 | 2.0 | 1.0 |
| 5 | Comum sistema de saúde fraco, de transporte deficiente e custo de vida excessivo, sofremos nesta terra de migalheira. | -3.0 | -3.0 | -4.0 |
| 6 | Os homens são animais muito estranhos, uma mistura de nervosismo de um cavalo, de teimosia de uma mula e malicia de um camelo | -3.0 | -1.6 | 2.0 |
| 7 | Parece-me absurdo que as leis são expressão da vontade pública , que abominam e punem o homicídio, o cometam elas mesmas e que para dissuadir o cidadão do assassínio, ordenem um assassínio público | -3.0 | -4.0 | -5.0 |
| 8 | Quando eu deixar a presidência, vou continuar a morar no meu apartamento, na mesma distância do sindicato que me protegeu da política, o que vai mais me dar orgulho é eu vou pode acordar de manhã e olhar para qualquer trabalhador e dizer para ele bom dia companheiro. | 3.0 | 2.6 | 4.0 |
| 9 | Temos todas as condições para nos afirmarmos enquanto escola europeia de formação de ativos, nomeadamente através de centros de formação profissionais de excelência de que dispomos apostando na competência e saber fazer, que os distingue os nossos profissionais | 3.0 | 5.0 | 4.0 |
| 10 | E eu sempre desejei isso para mim mesmo. | 1.0 | 0.0 | 0.0 |

Tabela 4.1: Previsão de sentimento em frases genéricas simples.

Quanto ao desempenho do sistema na classificação de frases genéricas simples, foi atribuída uma nota satisfatória. Entretanto, apesar disso, ficou o desafio de superar os constrangimentos encontrados na atrbuição de polaridades às palavras que não fazem partes das entradas da base de dados lexical utilizada.

Salientamos que, devido ao número elevado de frases analisadas (151), não foi possível listar todas elas. Por isso como ilustração, para cada categoria de experiências, apresentámos somente amostras de 10 frases. Os resultados quantitativos, sobre as 151 frases, encontra-se na

Secção 4.1.5.

### 4.1.2 Experiências com expressões idiomáticas genéricas

Para avaliar, o desempenho do sistema na previsão de sentimento em frases que contêm expressões idiomáticas genéricas, foi criado um dataset formado, na sua maior parte, de frases com expressões idiomáticas de Sentilex-Pt e OpLexicon[1]. Nesta categoria de dados, os resultados das experiências realizadas atingiu um nível que pode ser considerado como satisfatório. Portanto, apesar disso, o maior desafio está na deteção de sentimento em novas expressões idiomáticas. Isto é, tratar das construções gramaticais que, pelo sentido no qual são empregues nas frases, são expressões idiomáticas, mas a base de dados lexical utilizada não as apresenta nesta categoria. Como é o caso do Exemplo 5 da Tabela 4.2, *trazer à vida*, pode ser considerado como expressão idiomática e significa ressucitar ou curar.

| Nº | Frases | Pol. real | Pol. Sistema Sentilex-Pt | Pol. Sistema OpLexicon |
|---|---|---|---|---|
| 1 | Via a luz do fundo do túnel | 1.0 | 1.0 | 0.0 |
| 2 | Fazias vista grossa | -1.0 | -1.0 | 0.0 |
| 3 | Fizeste orelha mouchas | -1.0 | -1.0 | 0.0 |
| 4 | Dão com a línguas os dentes | -1.0 | -1.0 | 0.0 |
| 5 | Ele o trouxe à vida com o seu toque poético | 1.0 | 0.0 | 0.0 |
| 6 | Chamar o atenção de viajantes | -1.0 | .0.0 | -1.0 |
| 7 | O argumento apresentado dá sustentação aos ideias do primeiro presidente da república | 1.0 | 0.0 | 1.0 |
| 8 | Temos que saber dar valor à vida | 1.0 | 1.0 | 1.0 |
| 9 | O meu diretor é uma digna de confiança | 1.0 | 1.0 | 1.0 |
| 10 | O estudante estava a falar sozinho | 1.0 | 1.0 | 0.0 |

Tabela 4.2: Previsão de sentimento em frases com expressões idiomáticas genéricas.

### 4.1.3 Experiências com frases genéricas que contêm advérbios de intensidade.

Na realização das experiências com esta categoria de frases, depois de executarmos o sistema, nas 151 frases genéricas utilizadas, identificámos as que contém advérbios de intensidade. De seguida, em cada caso, procedemos a comparação da avaliação do sistema à avaliação humana (pré-definida no data set). O resultado desta observação levou-nos a classificar o nível do desempenho do sistema quanto à previsão de sentimento em frases deste género como satisfatório.

Na Secção 3.1.8, explicámos que a base de dados dos advérbios de intensidade, é composta de dois ficheiros de texto: o primeiro contem uma lista de atenuadores e o segundo uma lista de amplificadores. O fato da classe de advérbios em Português ser aberta (categoria gramatical na qual novas palavras são facilmente criadas) nos deixa o desafio de acrescentar sempre, em função das novas constatações, novos vocábulos nessas listas para aumentar o desempenho do sistema.

Ilustramos esses casos nas frases da Tabela 4.3, escolhidas como amostra das várias tratadas nas experiências realizadas com estas categorias de frases.

[1]Os detalhes sobres os dois léxicos encontra-se na Secção 2.1.4

| Nº | Frases | Pol. real | Pol. Sistema Sentilex-Pt | Pol. Sistema OpLexicon |
|---|---|---|---|---|
| 1 | A produção agrícola do país foi **muito** útil para alimentar o nosso povo. | 1.6 | 1.6 | 1.6 |
| 2 | Esta edição da feira do livro recebeu **menos** vistantes eloquentes | 0.6 | 0.6 | 0.6 |
| 3 | O meu médico realizou a cirurgia com métodos **mais** sofesticados | 1.6 | 1.6 | 1.6 |
| 4 | A estrategias adotadas foi **bastante** suficiente para convencer os investidores. | 2.6 | 2.6 | 2.6 |
| 5 | O país é rico em recursos naturais, mesmo economicamente é considerado como **bastante** pobre. | -0.6 | -0.6 | -1.6 |
| 6 | A camaleão é **muito** lento | -1.6 | -1.6 | -1.6 |
| 7 | As receitas arrecadadas neste perióodo **não** são **muito** úteis para apoiar a situação em questão. | 2.0 | 2.0 | 2.0 |
| 8 | O aluno com quoeficiente inteletual **mais** elevado do mundo é africano. | 2.0 | 1.6 | 1.6 |
| 9 | As mulhers daquela região são **maioritariamente** atenciosas | 2.0 | 1.6 | 1.6 |
| 10 | A lei de repatriamento de capitais foi votada no parlamento, por isso que os dirigentes dos partidos da oposição estão **tão** felizes. | 2.0 | 1.6 | 2.6 |

Tabela 4.3: Previsão de sentimento com frases contendo adverbios de intensidade.

### 4.1.4 Experiências com frases genéricas negativas

Apesar da complexidade da negação em Português, as regras propostas para o seu tratamento neste projeto ajudaram a alcançar resultados satisfatórios na avaliação do sistema quanto à previsão de sentimento nas frases afetas a este fenómeno linguístico.

Entretanto, salienta-se que a negação é um fenómeno linguístico muito complexo. Existem várias formas de negação na língua portuguesa; umas explícitas e outras implícitas. Este facto torna difícil o seu tratamento automático. A estratégia adotada neste trabalho para superar este obstáculo foi de criar, numa primeira fase, regras que abrangem os casos mais frequentes deste fenómeno gramatical, e progressivamente em função das novas constatações, essas regras serão atualizadas para abranger os casos mais complexos.

A Tabela 4.4 apresenta uma amostra do conjunto de frases negativas genéricas analisadas.

| Nº | Frases | Pol. real | Pol. Sistema Sentilex-Pt | Pol. Sistema OpLexicon |
|---|---|---|---|---|
| 1 | Acho que o Brexit **não** vai ser uma coisa maravilhosa para o vosso país e o Reino será desunido. | -2.0 | -2.0 | -2.0 |
| 2 | A situação política **não** está boa, por isso que o ministro do comercio foi para Benguela | -1.0 | -1.0 | -1.0 |
| 3 | **Nenhum** ser humano está feliz numa relação sem liberdade | -2.0 | -2.0 | -3.0 |
| 4 | Os jovens que organizaram a manifestação na província de Malanje **não** libertados. | -1.0 | -1.0 | -1.0 |
| 5 | Olha para o pico daquela montanha, **ninguém** é capaz de chagar a pé. | -1.0 | -1.0 | -1.0 |
| 6 | O povo angolano de natureza **não** é corrupto, o problema está nos governantes. | -2 | 0.0 | 0.0 |
| 7 | O relatório da Organização Mundial da Saúde sobre a situação sanitário na Libéria **não** é alarmante, como se esperava. | 1.0 | 1.0 | 1.0 |
| 8 | Os estudantes **não** querem ter aulas hoje, nem entrar no laboratório. | -2 | -1.0 | -1.0 |
| 9 | O perfume de origem magrebina **não** têm cheiro muito agradavel. | 1.0 | 1.0 | 1.0 |
| 10 | O Presidente angolano **não** vai reunir-se com o FMI em outubro e está prevista uma visitar da diretora geral do FMI a Angola em dezembro. | 0.0 | 0.0 | 0.0 |

Tabela 4.4: Previsão de sentimento em frases negativas

### 4.1.5 Aplicação de medidas de avaliação nas experiências com frases genéricas

No Processamento de Linguagem Natural, existem várias medidas que permitem quantificar o desempenho de sistemas. Para a classificação da performance do SentiSoft, foram utilizadas o Acerto e o Erro Absoluto Médio.

o Acerto é utilizado no domínio da recuperação de informação para medir a performance de sistemas na busca de documentos relevantes [MB94]. No seu contexto inicial, ela é definida da seguinte forma:

$$Acerto = \frac{Numerototaldedocumentosrelevantesrecuperados}{Numerototaldedocumentosrecuperados} \quad (4.1)$$

Utilizada no âmbito da AST, essa medida serviu de termómetro para medir o desempenho do sistema desenvolvido na previsão de sentimento. Assim sendo, o Acerto determinou a probabilidade do sistema em classificar corretamente os sentimentos.

Para permitir aplicação desta medida na avaliação do SentiSoft quanto à previsão de sentimento em texto genérico, foi feito uma observação dos resultados apresentados pelo sistema na previsão de sentimento das 151 utilizadas. A Tabela 4.5 apresenta em resumo os dados obtidos.

| | Avaliação Sentilex-Pt | Avaliação OpLexicon | Avaliação humana |
|---|---|---|---|
| **Número de frases avaliadas** | 151 | 151 | 151 |
| **Frases com sentimento positivo** | 61 | 72 | 74 |
| **Frases com sentimento negativo** | 32 | 26 | 38 |
| **Frases com sentimento neutro** | 58 | 53 | 39 |

Tabela 4.5: Dados estatísticos das experiências em texto genérico.

De seguida, foi feita a comparação do sentimento real ao atribuído pelo sistema para cada frase uma das 151 frases. Os diferentes resultados observados foram em 6 categorias diferentes:

a) True Positivo (TPos), frases com sentimento positivo corretamente classificadas;

b) True Negative (TNeg), frases com sentimento negativo corretamente polarizadas;

c) True Neutral (TNeu), frases com sentimento neutro corretamente classificadas;

d) False Positive (FPos), frases com sentimento positivo incorretamente avaliadas;

e) False Negative (FNeg), frases com sentimento negativo incorretamente classificadas;

f) False Neutral (FNeu), frases com sentimento neutro incorretamente polarizadas.

Baseando-se nas categorias acima referidas e em função do objetivo que se pretende alcançar, neste trabalho definimos o Acerto como a relação entre o número de frases corretamente classificadas e o número total de frases analisadas.

$$Acerto = \frac{TPos + TNeg + TNeu}{TPos + TNeg + TNeu + Fpos + Fneg + Fneu} \quad (4.2)$$

Assim sendo, para se obter um sistema de qualidade, espera-se que o número de frases corretamente polarizadas seja o maior possível. A Tabela 4.6 apresenta os dados obtidos com a repartição das frases de acordo com as suas categorias.

| | | | | | |
|---|---|---|---|---|---|
| **Sentilex-Pt** | | **Previsão humana** | | | |
| | **Previsão do sistema** | | **Positivo** | **Negativo** | **Neutro** |
| | | **Positivo** | 58 | | 3 |
| | | **Negativo** | 4 | 26 | 2 |
| | | **Neutro** | | 20 | 38 |
| **OpLexicon** | | **Previsão humana** | | | |
| | **Previsão do sistema** | | **Positivo** | **Negativo** | **Neutro** |
| | | **Positivo** | 53 | | 19 |
| | | **Negativo** | 3 | 18 | 5 |
| | | **Neutro** | | 12 | 41 |

Tabela 4.6: Matrizes confusão para as experiências em texto genérico.

Portanto, substituindo os valores da tabela acima na Fórmula (4.2), obtivemos o seguinte:

a) na base em léxico Sentlex-Pt

$$Total = 58 + 26 + 38 + 6 + 3 + 20 = 151 \quad (4.3)$$

$$Acerto = \frac{58 + 26 + 38}{151} = \frac{122}{151} = 0.807 \quad (4.4)$$

b) na base em léxico OPLexicon

$$Total = 53 + 19 + 8 + 18 + 12 + 41 = 151 \quad (4.5)$$

$$Acerto = \frac{53 + 18 + 41}{151} = \frac{112}{151} = 0.741 \quad (4.6)$$

Os resultados mostram que a probabilidade do sistema classificar corretamente o sentimento expresso em texto genérico é de 80.7% e de 74.1% na base em léxicos Sentilex-Pt e OpLexicon, respetivamente.

o Acerto limita-se somente na medição da performance considerando os casos corretamente classificados. No domínio de AST, essa análise é importante, mas também é claro que não é muito refinada, pois, não pormenoriza a margem de erro cometido no valor da polaridade. De acordo com [err18], desde o momento em que se calcula um resultado por aproximação, é preciso saber como estimar e delimitar o erro ali cometido. Para se entender melhor essa realidade, vamos utilizar de forma ilustrativa os exemplos apresentados na Tabela 4.7. Os dados

| Frase nº | Res. do sistema | Res. real | Diferença |
|---|---|---|---|
| 1 | 3.0 | 1.0 | 2.0 |
| 2 | -1.0 | -2.0 | 1.0 |
| 3 | 3.0 | -3.0 | 6.0 |
| 4 | 1.0 | -1.0 | 2.0 |
| 5 | 0.0 | 1.0 | 1.0 |

Tabela 4.7: Exemplos pa a margem de erro em polaridades.

acima mostram que, em termos de polaridade, o sistema avaliado acertou nas duas primeiras frases($Acerto = \frac{2}{5} = 0.4$). Na última coluna, é possível observar a disparidade entre os dois resultados para cada frase. A seguir, apresentamos algumas das informações reveladas por esses dados:

a) a melhor classificação é na frase 2, porque para além de acertar no sinal da polaridade, tem a menor margem de erro;

b) a pior classificação é a frase 3, por ter polaridade totalmente oposto ao valor real.

Em razão disso, neste trabalho, implementámos a seguinte fórmula:

$$EA_i = \frac{|S_i - R_i|}{2 * (DM + \frac{1}{2} * DM * FC)} \quad (4.7)$$

Esta fórmula permite calcular o erro cometido na classificação de sentimento expresso numa frase através da relação entre a polaridade real, a atribuída pelo sistema, as polaridades máxima e mínima nas frases analisadas. Esta relação depende dos sinais das polaridades atribuídas à frase (polaridade real e do sistema). Se forem iguais, menor será o erro. Assim sendo, nesta fórmula, $S_i$ representa a polaridade atribuída pelo sistema; $R_i$ a polaridade atribuída por avaliação humana; $DM$ (Diferênça Máxima) a diferença entre a polaridade máxima e mínima. Para o controlo de sinais de polaridades, foi introduzido o $FC$ (Fator de Controlo). O sinal deste último é positivo quando as duas polaridades são iguais, negativo quando são opostas e zero quando uma delas é zero.

$$FC = \left\{ \begin{array}{ll} 1, & \text{se } R_i * S_i > 0 \\ 0, & \text{se } R_i * S_i = 0 \\ -1, & \text{se } R_i * S_i < 0 \end{array} \right\} \quad (4.8)$$

Desta forma, a média aritmética dos Erros Absolutos de todas as frases corresponde ao Erro

Absoluto Médio (EAM).

$$EAM = \frac{1}{N}\sum_{i=1}^{N} E_i \quad (4.9)$$

A avaliação de um sistema de AST demonstra qualidade caso o Acerto seja elevado e o Erro Médio Absoluto baixo. Para os exemplos da Tabela 4.7, temos os seguintes resultados:

| Frases | Erro Absoluto | |
|---|---|---|
| 1 | 0.11 | 11% |
| 2 | 0.06 | 6% |
| 3 | 1.00 | 100% |
| 4 | 0.34 | 34% |
| 5 | 0.17 | 17% |

Tabela 4.8: Ilustração do Erro Absoluto Médio.

$$EAM = \frac{0.11 + 0.06 + 1 + 0.34 + 0.17}{5} = \frac{1.68}{5} = 0.33 \quad (4.10)$$

No caso dos resultados das experiências realizadas neste trabalho, existe um detalhe que merece destaque: os valores previstos e os encontrados estão em escalas diferentes, [-3,3] e [-6,6], respetivamente. Nesta situação, o Erro Absoluto Médio só é calculado depois da normalização desses intervalos. Assim sendo, os valores da escala [-6,6] foram convertidos para o intervalo [-3,3].

Deste modo, para cada uma das 151 frases genéricas analisadas, calculou-se o valor do Erro Absoluto e, de seguida, na base nos resultados obtidos, aplicou-se a Fórmula (4.9) para calcular o Erro Absoluto Médio. A Tabela 4.9 apresenta 5 casos, escolhidos como amostra.

| Nº | Frases | Pol. Sentilex-Pt | Pol. OpLexicon | Pol. real | Erro Absoluto | |
|---|---|---|---|---|---|---|
| | | | | | Sent. | OpLex. |
| 1 | É com particular emoção que saúdo com alegria todas as angolanas e angolanos, de Cabinda ao Cunene e do mar ao Leste. | 1.0 | 0.0 | 1.0 | 0.0% | 8.3% |
| 2 | Quanto à nossa defesa comum, rejeitamos como falsa a escolha entre a nossa segurança e os nossos ideais. | 1.0 | -2.0 | 1.0 | 0.0% | 50% |
| 3 | Sugere-se que o nosso dirigente não seja tolerante aos casos extremos. | -1.0 | -1.0 | -1.0 | 0.0% | 0.0% |
| 4 | Onde a resposta for sim, tencionamos seguir em frente. | 0.0 | 0.0 | 1.0 | 8.3% | 8.3% |
| 5 | E eu sempre desejei isso para eles. | 0.0 | 0.0 | 0.0 | 0.0% | 0.0% |

Tabela 4.9: Erro Absoluto em texto genérico.

A Tabela 4.10, a seguir, apresenta o Erro Absoluto Médio obtido para as experiências em texto genérico.

| | Erro Médio Absoluto | |
|---|---|---|
| **Sentilex-Pt** | 0.058 | 5.8% |
| **OpLexicon** | 0.076 | 7.6% |

Tabela 4.10: Erro Médio Absoluto em frases genéricas.

### 4.1.6 Breves considerações sobre as experiências com texto genérico

Resumidamente, é possível observar, através da Tabela 4.11, os diferentes resultados obtidos durante as experiências realizadas com texto genérico.

| | Sentilex-Pt | OpLexicon |
|---|---|---|
| **Total de frases analisadas** | 151 | 151 |
| **True Positive** | 58 | 53 |
| **True Negative** | 26 | 18 |
| **True Neutro** | 38 | 41 |
| **Taxa de acerto** | 80.7% | 74.7% |
| **Erro Absoluto Médio** | 5.8% | 7.6% |

Tabela 4.11: Resumo de resultados obtidos com texto genérico.

Os problemas relacionados com a flexibilidade da gramática portuguesa (surgimento de novos vocábulos, complexidade da negação, etc.) foram os principais obstáculos encontrados na automatização do processo de AST. Apesar disso, as medidas de avaliação provaram a eficácia do sistema no cumprimento dos requisitos pelos quais foi desenvolvido. A taxa de acerto esteve entre 74.1% e 80.7%. Da mesma forma a média da margem de erro cometido na atribuição de polaridades esteve abaixo de 10%. Tendo em conta esses dados, concluiu-se que o sistema desenvolvido é adequado para Análise de Sentimento em Texto.

## 4.2 Experiências com texto do domínio económico

Nesta secção, foram realizadas experiências para averiguar se os léxicos genéricos para AST em Português podem ser utilizados na Economia como alternativas para suprir o défice dessas ferramentas nesse domínio. Para tal, utilizámos as 64 frases do data set de frases do domínio ecónomico[2] como entrada e a base de dados lexical foi composta de Sentilex-Pt, OpLexcon e OpSentiLexicon. Este último é um léxico elaborado exclusivamente neste trabalho, pela fusão dos dois léxicos anteriores. Os detalhes sobre este dicionário de palavras com sentimento encontram-se na Subsecção 4.2.4. O foco das experiências foi as frases simples e as que contêm expressões idiomáticas.

### 4.2.1 Experiências com frases simples do domínio económico

Nesta secção, nas frases simples, para além daquelas que possuem as características definidas na Subsecção 4.1.1, acrescentaram-se também as frases negativas e as que contêm advérbios

[2]A descrição detalhada sobre o data set de frases do domínio económico encontra-se na Secção 4.1

de intensidade. A Tabela 12 apresenta algumas das várias frases, com as características acima referenciadas, utilizadas nesta subsecção.

| Nº | Frases | Pol. real | Pol. sistema Sentilex-Pt | Pol. sistema OpLexicon |
|---|---|---|---|---|
| 1 | **Perderam-se** casas, empregos foram *extintos*, negócios **encerraram**. | -3.0 | 0.0 | 2.0 |
| 2 | Um verdadeiro homem de negócios ou empreendedor não tem inimigos. | 3.0 | 3.0 | 1.0 |
| 3 | Os principais índices bolsistas dos EUA encerraram em **baixa** no dia em que regressaram à negociação depois do feriado do Memorial Day na segunda-feira. | -1 | 0.0 | 1 |
| 4 | A **greve** de camionista fez com o que o mercado não fosse abastecido com os produtos de primeira necessidade. | -1.0 | 0.0 | -1 |
| 5 | As bolsas europeias encerraram a marcar a maior **queda** desde Março. | -1.0 | 0.0 | 0.4 |
| 6 | A possibilidade de Itália precisar de ir de novo à urnas, no que poderá ser um teste à permanência do país da Zona Euro, tem sido uma das principais causas para o **receio** dos investidores. | 3.0 | 0.0 | -1.0 |
| 7 | Vamos **construir** estradas e pontes, redes eléctricas e linhas digitais que **alimentam** o nosso comércio e nos ligam uns aos outros. | 3.0 | 0.0 | -1.0 |
| 8 | O mercado de produtos alimentares neste país é muito competetitivo. | 2.0 | 0.0 | 1.0 |
| 9 | Se a conjuntura externa se degradar ? um risco que a turbulência em Itália veio evidenciar ? Portugal será significativamente **afectado** | -3.0 | -1.0 | 1.0 |
| 10 | O dinheiro público aplicado na educação é um **investimento** e não um gasto, pois ajuda a **construir** um futuro mais digno para as pessoas e para o país. | 3.0 | 3.6 | 1.6 |

Tabela 4.12: Amostra de frases simples do domínio economia.

Observando os dados da tabela acima, apesar de representarem uma amostra de várias frases simples (do domínio económico) analisadas, é notável que os resultados apresentados tanto por Sentilex-Pt, como por OpLexicon estão longe das expectativas. A análise detalhada de cada frase, junto com o seu resultado, aponta os motivos evidenciados no próximo parágrafo como principais causas deste insucesso.

O sentimento de uma palavra depende da sua especificação semântica no contexto em que este é empregue. A atrbuição de polaridade às entradas do Sentilex-Pt foi feita, especificamente, na base dos seus sentidos semânticos sobre entidades humanas [CP15]. Em vista disso, a sua aplicação no contexto económico apresenta piores resultados, essencialmente pelo seguinte:

a) A inexistência nesse léxico de vários vocábulos que no domínio em referência expressam sentimentos. Consequentemente, na classificação de frases contendo os termos em questão, mesmo expressando sentimento, por omissão o sistema os considera como neutros. Nos exemplos da Tabela 4.12, esses termos estão em negrito. O pior é que, em algumas frases, essas palavras são chave para a decisão do sentimento. Como é o caso de *queda* na Frase 5 (*as bolsas europeias encerraram a marcar a maior* ***queda.***)

b) A existência de vocábulos cujos valores sentimentais no léxico, são diferentes (até mesmo

opostos) dos que teriam no contexto em estudo. Temos os exemplos de *gordo* e *abastecido*, economicamente falando, o primeiro termo (*salário* ***gordo***) é sinónimo de *grande, enorme, abundante*. O segundo (*mercado* ***abastecido*** *com produtos de primeira necessidade*) significa *provido, munido, fornecido, dotado*, etc. Logo os dois termos, no domínio económico, expressam sentimento positivo. Porém, o Sentilex-Pt atribui-lhes polaridades -1 e 0, respetivamente.

### 4.2.2 Experiências com expressões idiomáticas da área económica

No conjunto das 64 frases, utilizadas nessas experiências, algumas delas contêm expressões idiomáticas típicas para a área económica. Identificámos cada uma elas, comparámos o resultado do sistema ao real (Resultado da avaliação humana).

À semelhança do que aconteceu com as frases simples, a classificação das expressões idiomáticas voltadas especificamente para a área económica, na base dos dois léxicos, apresentou dificuldades. O diagnóstico sobre os diferentes casos, apontou o seguinte como principal causa:

Em certo contexto, o sentimento expresso em palavra depende do argumento com que este se constrói. Esta situação é muito frequente nos adjetivos, estes mudam de polaridade em função do substantivo que (eles) modificam. O exemplo é do adjetivo *gelada* nas duas frases seguintes: *comida quente* e *cerveja quente*. Na primeira frase, *quente* expressa o sentimento positivo e na segunda, o mesmo adjetivo expressa o sentimento contrário, enquanto nos dois léxicos esse adjetivo tem polaridade fixa (zero). Semelhante a este, existem várias construções comunicativas multipalavra (substantivo+adjetivo, substantivo+preposição+substantivo, frases completas, etc.) que possuem significados próprios no domínio económico. Logicamente elas merecem o mesmo tratamento que as expressões idiomáticas, o que não acontece no Sentilex-Pt, muito menos no OpLexicon. No caso dos exemplos acima citados, em cada frase, o adjetivo *gelada* e o substantivo por ele modificado formam um par válido (substantivo+adjetivo) que deve receber uma polaridade única.

O desafio nesta temática consiste em detetar pares válidos, quando, nas frases, os seus componentes estão separados. Como ilustração, reformulamos os exemplos do parágrafo precedente sobre o adjetivo *gelada*.

a) A *cerveja* que tomámos, ontem depois do jantar, era *gelada*.

b) A *pizza* daqui sempre é entregue *gelada*.

Nestes exemplos, se não identificarmos os subtantivos que *gelada* qualifica, reduz-se significativamente a probabilidade de ter uma classificação correta nessas frases.

A Tabela 4.13 apresenta uma amostra de frases analizadas nesta etapa deste trabalho.

| Nº | Frases | Pol. real | Pol. sistema Sentilex-Pt | Pol. sistema OpLexicon |
|---|---|---|---|---|
| 1 | São Paulo a capital económica do Brasil, decretou hoje o **estado de emergência** devido ao excesso de protestos causados pela greve dos camionistas | -3.0 | -1.0 | 3.0 |
| 2 | **Um enfraquecimento do crescimento** da zona euro afectaria significativamente Portugal. | -3.0 | 0.0 | 0.0 |
| 3 | Se tem um governo que tem sido implacável no **combate à corrupção**, é o meu. | 1.0 | -2.0 | 1.0 |
| 4 | Vamos **recolocar a ciência no seu devido lugar** e **dominar as maravilhas** da tecnologia para elevar a qualidade do serviço de saúde e diminuir o seu custo. | 3.0 | 0.0 | 1.0 |
| 5 | Vamos **domar o sol** e os ventos e a terra para abastecer os nossos carros e **pôr a funcionaras** nossas fábricas. | 2.0 | 0.0 | 0.0 |
| 6 | O estado económico da empresa está num **crescimento anémico**, causado pelo **investimento reduzido**. | -2.0 | 0.0 | 0.0 |
| 7 | As reservas do fundo social devem esse ano crescer a **taxa ambiciosa**. | -2.0 | 0.0 | 1.0 |
| 8 | As expressões proferidas no discurso do ministro sobre a situação económica do pais são **verdades inconvenientes**. | -1.0 | 0.0 | 1.0 |
| 9 | A realidade é que ninguém quer contribuir para o **nevoeiro cerado** que se criou a volta da crise de divisa no país. | -2.0 | 0.0 | -1.0 |
| 10 | O ministro da inovação **deu luz verde** ao financiamento do projeto que lhe foi apresentado. | 1.0 | 1.0 | 0.0 |

Tabela 4.13: Amostra de frases com expressões idiomáticas da área económica.

Nos exemplos da tabela acima, podem ser consideradas como expressões idiomáticas do domínio económico, as seguintes construções multipalavra:
*excesso de protestos, combater a corrupção, recolocar a ciência no seu devido lugar, enfraquecimento do crescimento, dominar as maravilhas da tecnologia, por a funcionar, crescimento anémico, investimento reduzido, taxa ambiciosa, verdades inconvenientes, nevoeiro cerado, deu luz verde, etc.*

### 4.2.3 Aplicação de medidas de avaliação nas experiênças com frases do domínio económico

Resumidamente, a Tabela 4.14 apresenta os dados os obtidos nas experiências realizadas com o texto do domínio económico.

| | Avaliação na base em Sentilex-Pt | Avaliação na base em OpLexicon | Avaliação humana |
|---|---|---|---|
| **Número de frases avaliadas** | 64 | 64 | 64 |
| **Frases com sentimento positivo** | 18 | 38 | 23 |
| **Frases com sentimento negativo** | 12 | 13 | 36 |
| **Frases com sentimento neutro** | 34 | 13 | 5 |

Tabela 4.14: Dados estatísticos da 64 frases analisadas na base de Sentilex-Pt e OpLexCon.

A partir dos dados da Tabela 4.14 e os obtidos comparando o resultados da avaliação humana ao do sistema para cada uma das 64 frases do dataset, obtivemos as matrizes confusão representadas através da Tabelas 4.15.

| | | | | | |
|---|---|---|---|---|---|
| **Sentilex-Pt** | | **Avaliação humana** | | | |
| | **Previsão do sistema** | | **Positivo** | **Negativo** | **Neutro** |
| | | **Positivo** | 11 | | 7 |
| | | **Negativo** | 2 | 9 | 1 |
| | | **Neutro** | | 32 | 2 |
| **OpLexicon** | | **Avaliação humana** | | | |
| | **Previsão do sistema** | | **Positivo** | **Negativo** | **Neutro** |
| | | **Positivo** | 13 | | 25 |
| | | **Negativo** | 4 | 6 | 3 |
| | | **Neutro** | | 12 | 1 |

Tabela 4.15: Matrizes confusão para as 64 frases do domínio económico.

Aplicando as Fórmulas (4.2) e (4.9), obtivemos:
a) na base em léxico Sentilex-Pt

$$Total = 11 + 9 + 2 + 7 + 32 + 3 = 64 \quad \textbf{(4.11)}$$

$$Acerto = \frac{11 + 9 + 2}{64} = \frac{22}{64} = 0.34 \quad \textbf{(4.12)}$$

$$EAM = 0.15 \quad \textbf{(4.13)}$$

b) na base em léxico OpLexicon

$$Total = 64 \quad \textbf{(4.14)}$$

$$Acerto = \frac{13 + 6 + 1}{64} = \frac{20}{64} = 0.31 \quad \textbf{(4.15)}$$

$$EAM = 0.19 \quad \textbf{(4.16)}$$

No total das 64 frases do dataset, o Sentilex-Pt teve uma taxa de acerto de 34%, o que lhe coloca numa posição de ligeira vantagem em relação ao OpLexicon que teve somente 31%. No entanto, esses resultados demonstram que os dois léxicos não são adequados para Análise de Sentimento em Texto no domínio económico.

### 4.2.4 Experiências na base em léxico OpSentiLexicon

Já que as experiências na base nos dois léxicos anteriores, isoladamente, não tiveram sucesso, com o intuito de elevar a taxa de acerto ao um nível minimamente aceitável, neste trabalho elaborou-se um novo léxico denominado *OpSentiLexicon*. Este dicionário de palavras com sentimento surgiu na fusão do Sentilex-Pt e OpLexicon.

$$OpSentiLexicon = SentilexPt + OpLexicon \quad \textbf{(4.17)}$$

No processo do desenvolvimento deste novo recurso para AST, para desambiguar os casos dos vocábulos comuns aos dois léxicos e que em cada um deles possuem polaridades diferentes, a

prioridade foi dada ao valor sentimental atribuído por Sentilex-Pt, pelo fato deste ser o léxico que apresentou melhor classificação nas experiências anteriores. O OpSentilexCon é composto de 94 210 formas flexionais, 34 929 das quais são expressões idiomáticas.

Realizaram-se experiências na base neste novo léxico com as 64 frases do domínio económico. Os resultados estão apresentados resumidamente através das Tabelas 4.16 e 4.17.

| | **Avaliação na base em OpSentiLexicon** | **Avaliação humana** |
|---|---|---|
| **Número de frases avaliadas** | 64 | 64 |
| **Frases com sentimento positivo** | 37 | 23 |
| **Frases com sentimento negativo** | 15 | 36 |
| **Frases com sentimento neutro** | 12 | 5 |

Tabela 4.16: Dados da avaliação da aplicação de OpSentilexCon em AST do domínio económico.

| | | **Previsão humana** | | |
|---|---|---|---|---|
| | | **Positivo** | **Negativo** | **Neutro** |
| **Previsão do sistema** | **Positivo** | 15 | | 22 |
| | **Negativo** | 2 | 8 | 5 |
| | **Neutro** | | 12 | 0 |

Tabela 4.17: Matriz confusão para experiência na base em OpSentiLexicon.

$$Acerto = \frac{15+8+0}{64} = \frac{23}{64} = 0.359 = 36\% \quad \textbf{(4.18)}$$

O OpSentilexiCon, aplicado em texto económico, apresentou uma taxa de acerto de 36%, melhora um pouquinho em relação aos dois outros léxicos. Ainda assim, esse resultado, não atinge um nível satisfatório para a utilização deste recurso em AST no domínio económico.

## 4.3 Considerações

Observando os dados da Tabela 4.18 e da Figura 4.1 é possível notar a redução considerável do desempenho dos léxicos genéricos quando são adaptados no domínio económico.

| | **Taxa de acerto** | |
|---|---|---|
| | **Texto genércio** | **Texto da Economia** |
| **Sentilex-Pt** | 80.7% | 34% |
| **OpLexicon** | 74.7% | 31% |
| **OpSentiLexicon** | ? | 36% |

Tabela 4.18: Desempenho dos léxicos gerais no domínio económico.

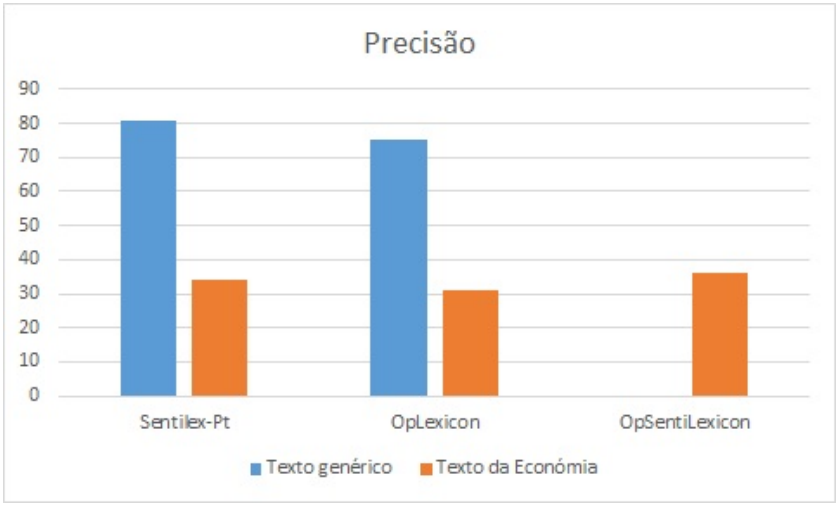


Figura 4.1: Desempenho dos léxicos gerais no domínio económico.

[KB13] e [Bel17] defendem a opinião de que o fator contexto tem influência na classificação de texto. A experiência adquirida neste trabalho reforça esta teoria mostrando que, quando se trata do contexto económico, esta influência é tão significativa que um léxico genérico não pode ser utilizado como alternativa, em caso de carência. Por conseguinte, a solução adequada para suprir a necessidade atual de classificar o texto do domínio económico em Português passa pela elaboração de um léxico específico a essa área. O novo léxico pode ser a versão económica de Sentilex-Pt. Para tal, sugerimos as seguintes alterações na versão atual.

a) Atualizar as polaridades dos vocábulos, sobretudo adjetivos e verbos, baseando-se nas especificações sintático-semânticas destes termos no contexto económico.

b) Acrescentar novas palavras com sentimento, polarizadas no sentido semântico sobre entidades do domínio económico.

c) Acrescenter às expressões idiomáticas atuais, estruturas multi-palavras especificas à Economia.

d) Alargar a escala de polaridades, permitindo assim acrescentar intensidades aos sentimentos.

O Anexo A.1 desta dissertação, apresenta uma proposta de termos e expressões idiomáticas, próprios para o contexto económico, que poderão ser proveitosos na elaboração deste léxico.

## 4.4 Sumário

Este capítulo é fundamental para dar resposta ao problema tratado neste trabalho. Numa primeira fase, avaliou-se o desempenho do SentiSoft. Na segunda, utilizou-se esta ferramenta para medir a capacidade dos léxicos genéricos em Português quanto à classificação de texto do domínio económico. No próximo capítulo, seguem as considerações finais e as propostas para o futuro trabalho nesta linha de investigação.

# Capítulo 5

# Conclusões e Trabalho futuro

Por fim, ultimamos esta dissertação com as considerações finais do trabalho, bem como as sugestões de novas abordagens para pesquisas futuras nesta linha de investigação que serão úteis para o bem da ciência e da sociedade, em geral. Assim sendo, apresentamos, de seguida, as principais conclusões desta dissertação e trabalho futuro que pode ser seguido.

## 5.1 Conclusões

O rápido crescimento de documentos textuais fez com que as organizações perdessem o controlo da informação sobre o ponto de vista dos seus clientes, quanto à qualidade de produtos e serviços por estes prestados. A Análise de Sentimento em Texto surgiu como solução a este problema, esta área do PLN se preocupa com o desenvolvimento de sistemas inteligentes capazes de determinar automaticamente os sentimentos em grandes quantidades de texto não estruturado. Esta tarefa é realizada com base em duas abordagens principais: a abordagem baseada em Aprendizagem Automática e a sustentada por léxico. Esta última é a abordagem na qual esta dissertação foi totalmente baseada.

Por ser uma área relativamente recente e pelo fato da maior parte de investigadores estarem dedicados somente ao desenvolvimento de ferramentas para o Inglês, este campo do PLN carece de recursos para o Português. Com o intuito de suprir o défice nesta língua, em léxico de Análise de Sentimento em Texto no domínio económico, neste trabalho pretendeu-se verificar se os seus léxicos genéricos apresentam resultados satisfatórios quando são adaptados neste domínio. A organização do trabalho baseou-se na proposta de [DdS05], na qual os trabalhos de PLN estão repartidos em três níveis: nível linguístico, linguístico-computacional e computacional. O nível linguístico preocupou-se dos aspetos sintáticos, semânticos, morfológico e pragmáticos do processo de AST: o pré-processamento de texto, a etiquetagem morfológica, tratamento de expressões idiomáticas, negação, ironia, etc. O nível linguístico-computacional foi responsável pela representação, numa linguagem tratável pelo computador, das diferentes propostas adotadas para dar solução ao problema em estudo e a última categoria encarregou-se no desenvolvimento do SentiSoft, sistema informático para AST.

A avaliação do desempenho do SentiSoft, em texto genérico, apresentou resultados satisfatórios com taxa de acerto de 80.7% e 74.7% na base em léxico Sentilex-Pt e OpLexicon, respetivamente. As experiências realizadas com o mesmo sistema em texto do domínio económico, na base nos dois léxicos acima referênciados, reduziram a taxa de acerto, para 34% e 31%. Esses dados provaram que esses recursos não produzem bons resultados quando são utilizados no domínio económico. A análise detalhada dos diferentes resultados obtidos apontaram a variação do sentido semântico dos vocábulos em função do contexto como principal causa deste insucesso.

Com a preocupação de encontrar mecanismos que melhorem esta classificação, elaborou-se, neste trabalho, fundido Sentilex-Pt e OpLexicon, o léxico OpSentiLexicon. O desempenho deste novo recurso para AST, no domínio económico, foi avaliado com 36% de taxa de acerto. Perante esta realidade, concluíu-se que os léxicos genéricos para AST em Português não podem ser utilizados no domínio económico para suprir com a carência deste recurso nesta área.

## 5.2 Trabalho futuro

Em função das experiências adquiridas ao longo deste trabalho de investigação, para darmos solução ao problema de carência de recursos de Análise Sentimento em Texto no domínio económico, sugerimos para os trabalhos futuros, a elaboração de um léxico específico voltado para esta área. Na Secção 4.3, propusemos uma abordagem que poderá ser proveitosa na elaboração deste léxico, assim como uma lista de palavras e expressões idiomáticas, específicas para a área económica, que poderão integrar o léxico sugerido.

Outra abordagem que poderá ser explorada, futuramente, nesta linha de investigação, é o recurso às técnicas de Aprendizagem Automática, treinando o sistema com um conjunto de palavras (frases) previamente polarizadas.

Ao longo desta dissertação, para dar solução às situações relacionadas com a complexidade da língua no tratamento automático de texto, criámos regras específicas para cada uma delas. Apesar dessas regras terem contribuído muito no desenvolvimento de um sistema para AST com um nível de desempenho satisfatório, acreditamos que elas não são completas. Para aumentar a eficácia do sistema, sugerimos, melhoramento dessas regras, abrangendo o maior número de casos possível.

Este trabalho deixa disponível para a comunidade dois data sets de frases com sentimento previamente atribuido, um composto de 151 frases genéricas e outro de 64 frases do domínio económico.

# Bibliografia


[ An06a] Ana Paula de Araújo. Advérbios de intensidade [online]. 2006. Available from: www.infoescola.com/portugues/adverbios-de-intensidade/ [cited 2018-02-15]. 16

[ An06b] Ana Paula de Araújo. Advérbios de negação [online]. 2006. Available from: https://www.infoescola.com/portugues/adverbios-de-negacao/ [cited 2018-02-12]. 14

[AC12] Shubhamoy Dey Anuj Charma. A comparative study of feature selection and machine learning techniques for sentiment analysis. Proceedings of the 2012 ACM Research in Applied Computation Symposium, pages 1–7, 2012. 21

[AE06] Fabrizio Sebastian Andrea Esuli. Sentiwordnet: A publicly available lexical resource for opinion mining. In In Proceedings of the 5th Conference on Language Resources and Evaluation, pages 417–422. In Proceedings of the 5th Conference on Language Resources and Evaluation, 2006. 13, 29

[AGLdS00] Viviane Moreira Aline Graciela Lermen dos Santos, Karin Becker. Um estudo de caso de mineração de emoções em textos multilíngues. Universidade Federal do Rio Grande do Sul, 0000. 11, 17

[AM12] J.L. Castro J.M. Zurita A. Moreo, M. Romero. Lexicon-based comments-oriented news sentiment analyzer system. Expert Systems with Applications, 39(10):9166–9180, 2012. 18

[AMR14] L.A. Ureña-López A. Montejo-Ráez, M.C. Díaz-Galiano. Crowd explicit sentiment analysis. Knowledge-based system, 69:134–139, 2014. 24

[AO13] Rosa M. Carro Alvaro Ortigosa, José M. Martín. Sentiment analysis in facebook and its application to e-learning. Computers in Human Behavior, 11:527–541, 2013. 10, 22

[Bel17] Francini Scipioni Belau. Uma proposta de representação linguístico-computacional da negação com vistas à análise de sentimentos em contexto de ensino e aprendizagem on-line. RDBU Repositório Digital da Bibilioteca Unismo, 42, 2017. 2, 9, 24, 51

[Cam13] Erik Cambria. An introduction to concept-level sentiment analysis. In Advaces in soft computing and its applications, volume 8266, page 478–483. Mexican international conference on artificial inteligence, 2013. 23

[CL09] Yulan He Chenghua Lin. Joint sentiment/topic model for sentiment analysis. AGM library, pages 375–384, 2009. 9, 21

[CP15] Silva Mário J. Carvalho Paula. Sentilex-pt: Principais características e potencialidades. Oslo Studies in Language, 7:425–438, 2015. xiii, 3, 9, 10, 46

[CS07] R. Mihalcea C. Strapparava. Semeval-2007 task 14: affective text reseachgate. In SemEval-2007 task 14: affective text Reseachgate, volume 10, pages 70–74. SemEval '07 Proceedings of the 4th International Workshop on Semantic Evaluations, 2007. 19

[CS14] R. Mihalcea C. Strapparava. Time corpora: Epochs, opinions and changes. Knowledge-based system, 69:3–13, 2014. 19

[Dat18] Data Cience Academy. Aprendizado de máquina [online]. 2018. Available from: https://pt.wikipedia.org/wiki/Aprendizado_de_maquina [cited 2018-04-13]. 8

[DdS05] Bento Carlos Dias-da Silva. Wordnet.br: An exercise of human language technology research. In Gwc 2006: Third International Wordnet Conference, Proceedings, pages 301–303. Masaryk University, 2005. 10, 53

[dSG16] Nuno Ricardo Pinheiro da Silva Guimarães. Lexicon-Expansio-System for Domain and Time Oriented Sentiment Analysis. Universidade de Porto, 2016. 20

[DT11] Isabel Azevedo Diogo Texeira. Análise de opiniões expressas nas redes sociais. Revista ibérica de sistemas e tecnologias de informação, (8):53–65, 2011. 18

[DZ15] Zengcai Su Yunfeng Xu Dongwen Zhang, Hua Xu. Chinese comments sentiment classification based on word2vec and svmperf. Expert Systems and aplcations, 42:1857–1863, 2015. 22

[EAMM07] Ana Paula L. Ambrósio Edison Andrade Martins Morais. Mineração de Textos. Universidade Federal de Goiás, 2007. 11

[err18] Teoria dos erros [online]. 2018. Available from: https://pt.wikipedia.org/wiki/Teoria_dos_erros [cited 2018-06-21]. 43

[Fel13] Ronen Feldman. Techniques and applications for sentiment analysis. communications of the acm, 56(4):82–89, 2013. 1

[For15] Ana Catarina Barbosa Forte. Análise de comentários de clientes com o auxílio a técnicas de Text Mining para determinar o nível de (in)satisfação. Universidade de porto, 2015. 19

[GW14] Jian Ma Kaiquan Xue-Jibao Gud Gang Wang, Jianshan Sun. Sentiment classification: The contribution of ensemble learning. Decision Suport System, 57:77–93, 2014. 22

[Has16] MiriamFernandeza HarithAlania HassanSaifa, YulanHeb. Contextual semantics for sentiment analysis of twitter. Information processing and management, 37:5–19, 2016. 20

[HL04] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In KDD 2004 Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 168–177. University of Illinois at Chicago, 2004. 23

[hul18] hultig [online]. 2018. Available from: http://hultig.di.ubi.pt/ [cited 2018-05-19]. 33

[IZ17] Abdelhak Lakhouaja Imad Zeroual. Arabic information retrieval: Stemming or lemmatization? In Intelligent Systems and Computer Vision. IEEE, 2017. 20

[JG96] Henry McGilton James Gosling. The Java Language Environment. Sun Microsystems, Inc., 1996. 33

[JL09] Stephanie Seneff Jingjing Liu. Review sentiment scoring via a parse-and-paraphrase paradigm. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, volume 1, pages 161–169. Association for Computational Linguistics, 2009. 15, 16

[JNO99] Luís Alfredo Amaral João Nuno Oliveira. O papel da qualidade da informação nos sistemas de informação. In Conferência Especializada em Sistemas E Tecnologias de Informação. Universidade do Minho, 1999. 3

[JP07] M. Francis J. Pennebaker, R. Booth. Linguistic inquiry and word count. libraries and learning services, 2007. 23

[Kau16] Anderson Uilian Kauer. Abordagem linguística na classificação automática de textos em portugês. UniversidaeE Federal do Rio Grande Do Sul, 2016. Available from: http://www.lume.ufrgs.br/bitstream/handle/10183/140910/000991520.pdf?sequence=1. 11, 22

[KB13] Diego Tumitan Karin Becker. Introdução à mineração de opiniões: Conceitos, aplicações e desafios. In Lectures of the 28th Brazilian Symposium on Databases. Universidade Federal do Rio Grande do Sul (UFRGS), 2013. 2, 51

[KC09] Dirk Van den Poel Kristof Coussement. Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. Expert Systems with Applications, page 6127–6134, 2009. 23

[KdS17] Daniel Dalip Karine de Souza, Moisés Pereira. Unilex: Método léxico para análise de sentimentos textuais sobre conteúdo de tweets em português brasileiro. Abakós, 5(2):79–96, 2017. 20

[KML13] Shamanth Kumar, Fred Morstatter, and Huan Liu. Twitter Data Analytics. Arizona State University, 2013. 10

[LD16] Anuraag Biswas Beepa Bose-Sweta Tiwari Lopamudra Dey, Sanjay Chakraborty. Sentiment analysis of review datasets using naïve bayes' and k-nn classifier. MECS, 8:54–62, 2016. 23

[Liu07] Bing Liu. Sentiment Analysis and Opinion Mining. Morgan Claypool, 2007. 2

[MAC95] Andre Vellino Michael A. Covington, Donald Nute. Prolog Programming in Depth. The University of Georgia, 1995. 7

[MB94] Fredric Gey Michael Buckland. The relationship between recall and precision. Journal of the ameircan society for information science, 12, 1994. 41

[MKS10] Wesam Ashour Motaz K. Saad. Arabic morphological tools for text mining. In Reseachgate. International Conference on Electrical and Computer Systems (EECS'10), 2010. 21

[MT11] Milan Tofiloski Kimberly Voll-Manfred Stede Maite Taboada, Julian Brooke. Lexicon-based methods for sentiment analysis. Association for Computational Linguistics, 37(2):267–307, 2011. 18

[MWDK10] Benjamin Roth Michael Wiegand, Alexandra Balahur and Andres Montoyo Dietrich Klakow. A survey on the role of negation in sentiment analysis. In Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pages 60–68. Association for Computational Linguistics, 2010. 15, 23

[Nie14] Finn ˚Arup Nielsen. A new anew: Evaluation of a word list for sentiment analysis in microblogs. Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages, pages 93–98, 2014. 12, 19

[NRS00] Flávia Barros Nelson Rocha Silva, Diego Lima. Sapair: Um processo de análise de sentimento no nível de característica. Centro de Informática, UFPE, Brasil, pages 1–10, 0000. 17

[Oli15] Daniel José Silva Oliveira. Avaliação de métodos de análise de sentimento em mídias socias na gestão social e política. Universidade Federal de Lavras, 2015. 10, 24

[PC09] Mário J. Silva Eugénio de Oliveira Paula Carvalho, Luís Sarmento. Clues for detecting irony in user-generated contents: oh...!! it's "so easy". In Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion, pages 53–56. ACM, 2009. 16, 17

[PC15] Erik Cambria Prerna Chikersal, Soujanya Poria. Sentu: Sentiment analysis of tweets by combining a rule-based classifier with supervised learning. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), page 647–651. Nanyang Technological University, 2015. 22

[Pe18] 2003-2018 Porto editora. frase in dicionário infopédia da língua portuguesa [online]. 2018. Available from: https://www.infopedia.pt/dicionarios/lingua-portuguesa/ [cited 2018-05-08 14:53:02]. 29

[PL06] Zaenen A. Polanyi L. Contextual valence shifters. In Computing Attitude and Affect in Text: Theory and Applications, volume 20, pages 1–10. Springer, Dordrecht, 2006. 15, 16, 30

[PM09] Richard D. Lawrence Prem Melville, Wojciech Gryc. Sentiment analysis of blogs by combining lexical knowledge with text classification. In international conference on Knowledge discovery and data mining, pages 1275–1282. international conference on Knowledge discovery and data mining, 2009. 18

[RAA14] Safa Bani Essa RMahmoud Al-Ayyoub. Lexicon-based sentiment analysis of arabic tweets. Internacional Jornal Social Network Mining, pages 1–15, 2014. 20

[Reg16] Reginaldo J. Santos. Portugueses passam diariamente mais de hora e meia nas redes sociais [online]. 2016. Available from: http://expresso.sapo.pt/sociedade/2016-09-21-Portugueses-passam-diariamente-mais-de-hora-e-meia-nas-redes-sociais [cited 2018-04-10]. 1

[SBS00] Andrea Esuli Stefano Baccianella and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. Consiglio Nazionale delle Ricerchel, pages 2200–2204, 0000. 18

[Sil15] Lucas Lo Ami Alvino Silva. Análise de sentimentos em contexto: estudo de caso em blog de empreendedorismo. Universidade de Brasília, 2015. 37

[SK14] Saif M. Mohammad Svetlana Kiritchenko, Xiaodan Zhu. Sentiment analysis of short informal texts. Journal of Artificial Intelligence Research, (50):723–762, 2014. 19

[SM10] P. Turney S. Mohammad. Emotions evoked by common words and phrases. In Proceedings of the NAACL HLT 2010 Workshop on Computational, page 26–34. Association for Computational Linguistics, 2010. 19

[SML15] Jiun-Hung Chen Shuhua Monica Liu. A multi-label classification based approach for sentiment classification. Expert systems with aplications, 42:1083–1093, 2015. 19

[SP12] Erik Cambria Peipei Yang Amir Hussain Tariq Durrani Soujanya Poria, Alexander Gelbukh. Merging senticnet and wordnet-affect emotion lists for sentiment analysis. In 2012 IEEE 11th International conference on Signal processing(ICSP). IEEE, 2012. 18

[ST08] Jin Zhang Songbo Tan. An empirical study of sentiment analysis for chinese documents. Expert Systems with Applications, 34:2622–2629, 2008. 2, 9, 20

[ST09] Yuefen Wang Hongbo Xu Songbo Tan, Xueqi Cheng. Adapting naive bayes to domain adaptation for sentiment analysis. In Advences in information retrievial, pages 337–349. Europe Conference of Information Retrievial, 2009. 21

[Sua00] Sua pesquisa.com. A revoluçaõ industrial [online]. 2000. Available from: `https://www.suapesquisa.com/industrial/` [cited 2018-02-18]. 3

[SW11] X. Song Y. Wei H. Li S. Wang, D. Li. A feature selection method based on improved fisher's discriminant ratio for text sentiment classification. Expert Syst. Appl., 38(7):8696–8702, 2011. 21

[Tab16] Maite Taboada. Sentiment analysis: Na overview from linguistics. Annual Review of Linguistics, 2:325–347, 2016. 13

[TH02] Ronald L. RIVEST Clifford STEIN Thomás H.CORMEN, Charles E. LEISERSON. Algoritmos, tradução da segunda edição americana. Editora Campus, 2002. 27

[TMSA14] Pyry Takala, Pekka Malo, Ankur Sinha, and Oskar Ahlgren. Gold-standard for topic-specific sentiment analysis of economic texts. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 168–177. European Language Resources Association (ELRA), 2014. 24

[TN15] Jeonghee Yi Tetsuya Nasukawa. Sentiment analysis: Capturing favorability using natural language processing. In Proceedings of the 2nd International Conference on Knowledge Capture, LNCS, 2015. 19

[VF11] Mantovanini unisalesiano Viviane Ferreira. a influência da avaliação de desempenho no desenvolvimento profissional. Centro Universitário Católico Salesiano, 2011. 37

[VH00] Janyce Wiebe Vasileios Hatzivassiloglou. Efects of adjective orientation and gradability on sentence subjectivity. In COLING: Proceedings of the 18th conference on Computational linguistics, volume 1, pages 299–305, 2000. 12

[VN13] Arjun Bhatia Vivek Narayanan, Ishan Arora. Fast and accurate sentiment classification using an enhanced naive bayes model. Intelligent Data Engineering and Automated Learning IDEAL, 8206, 2013. 22

[wik17] wikipedia. Processo de desenvolvimento de software [online]. 2017. Available from: https://pt.wikipedia.org/wiki/Processo_de_desenvolvimento_de_software [cited 2018-05-19]. 32

[Wik18] Wikipedia. Expressões idiomáticas [online]. 2018. Available from: https://pt.wikipedia.org/wiki/Expressõesidiomática [cited 08 Abril 2018]. 14

[YW00] Cheng-ming GUO James E. MCDONALD Tony PLATE Brian M. SLATOR Yorick WILKS, Dan FASS. Machine tractable dictionaries as tools and resources for natural language processing. Computing Research Laboratory, pages 750–755, 0000. 9

# Apêndice A

# Anexos

## A.1 Propota de palavras e expressões idomáticas da área económica

| Expressões polares | Exemplo | Sentimento |
|---|---|---|
| alimentar | Vamos construir estradas e pontes, redes eléctricas e linhas digitais que alimentam o nosso comércio e nos ligam uns aos outros. | positivo |
| apreender | A declaração de estado de emergência permite ao município fazer compras sem licitação ou apreender bens privados | negativo |
| baixa | Os principais índices bolsistas dos EUA encerraram em baixa. | negativo |
| combater à corrupção | O meu governo esta a evidenciar esforço no combate à corrupção, desde o primeiro dia. | positivo |
| construir | Vamos construir estradas e pontes, redes eléctricas e linhas digitais que alimentam o nosso comércio e nos ligam uns aos outros. | positivo |
| criar novos empregos | O estado da economia pede acção, corajosa e rápida, e nós vamos agir para criar novos empregos. | positivo |
| desenvolvimento | Aliado a todos os demais fatores, que já salientei, Portugal tem de criar uma matriz estratégica de desenvolvimento, da qual fazem parte os nossos jovens. | positivo |
| desestabilizado | A incerteza do panorama político italiano, que tem desestabilizado os mercados europeus | positivo |
| enfraquecimento do crescimento | Um enfraquecimento do crescimento da zona euro afectaria significativamente Portugal. | negativo |
| estado de calamidade pública | estado de emergência pode evoluir para um estado de calamidade pública. | negativo |
| falta de abstecimento | A câmara municipal advertiu, num comunicado, que caso a greve se prolongue, poderá ser declarado um dia de feriado nacional devido à falta de abastecimento | negativo |
| impacto | O facto de a Finlândia passar a cobrar impostos sobre as pensões que paga a esses reformados a viver em Portugal terá um impacto entre três a seis milhões de euros por ano. | negativo |
| licitação | A declaração de estado de emergência permite ao município fazer compras sem licitação | positivo |
| paralisação | Neste quinto dia de paralisação, que afecta os 27 estados brasileiros | negativo |
| pôr a funcionar | Vamos pôr a funcionar as nossas fábricas. | positivo |
| rasgar o acordo fiscal | O Parlamento finlandês aprovou a proposta do governo de rasgar o acordo fiscal com Portugal. | negativo |
| usufruido | Os reformados irlandês te usufruido de um estatuto de residêntes não habituais em Portugal. | positivo |

Tabela A.1: Lista de expressões polares económicas

## A.2 Extrato do dataset de frases genéricas

| Frases | Pol. |
|---|---|
| É com particluar emoção que saúdo com alegria todas as angolanas e angolanos, de Cabinda ao Cunene e do mar ao Leste. | 1 |
| Por nós, eles combateram e morreram, em lugares como Concord e Gettysburg. | 0 |
| Esta é a viagem que hoje continuamos. | 0 |
| Permanecemos o pais mais poderoso e próspero na Terra. | 2 |
| Estas coisas são verdadeiras. 1 | |
| Com o sistema de saúde fraco, de transporte deficiente e custo de vida excessivo, sofremos nesta terra de migalheira, enquanto isso, os governantes gozam com o povo. | -3 |
| E por que eu a abandonei? | -1 |
| Razão para que se argumente que seria mais relevante dirigir a mensagem a mulheres fortes e autónomas. | 3 |
| Esta é a fonte da nossa confiança ? | 1 |
| Temos todas as condições para nos afirmarmos enquanto escola europeia de formação de ativos, nomeadamente através dos centros de formação profissional de excelência de que dispomos, apostando na competência e no saber fazer, que distingue os nossos profissionais. | 3 |
| Quarenta e quatro americanos fizeram até agora o juramento presidencial. | 0 |
| No entanto, muitas vezes a tomada de posse ocorre no meio de nuvens espessas e furiosas tempestades. | -1 |
| Ao reparamos essa grandiosa nação, compreendemos que a grandeza nunca é um dado, é adquirido, é conquistada. | 3 |
| Woz e eu começamos a Apple na garagem dos meus pais quando eu tinha 20 anos. | 0 |
| Tenho de agradecer a todos os meus fãs, no mundo todo, por essa façanha. | 1 |
| Cantarei as canções que meus fãs gostam de ouvir. | 1 |
| O próximo álbum tem que ser três vezes melhor, não pode ser apenas bom, por que isso seria decepcionante, então tirarei meu tempo e tentarei fazer a coisa certa. | 3 |
| Agora que o dia está quase no seu fim, pense nos problemas, nas dificuldades, no estresse! | -2 |
| O México aproveitou-se dos EUA durante demasiado tempo, enormes défices comerciais e a pouca ajuda na muito débil fronteira, isso deve mudar, AGORA. | -3 |
| Acho que o Brexit vai ser uma coisa maravilhosa para o vosso país e um Reino Unido livre e independente é uma bênção para o mundo. | 3 |

Tabela A.2: Extrato do dataset de frases genéricas

## A.3 Extrato do dataset de frases da área económica

| Frases | Pol. |
|---|---|
| Finanças cativaram 611,5 milhões de euros de despesa até Março Valor das cativações no final do primeiro trimestre representa menos 377 milhões do que em igual período de 2017. | -1 |
| Assim, pelo menos, não acabaríamos numa situação em que não há acordo fiscal entre a Finlândia e Portugal. | 1 |
| O aviso é do Fundo Monetário Internacional, que até acredita que Portugal vai crescer ao ritmo prometido pelo Governo, mas que defende que o Executivo deveria antecipar o esforço de ajustamento face ao que tem programado. | 2 |
| Uma economia aberta cuja presença das exportações tem aumentado de forma bem-sucedida nos mercados, lê-se no comunicado da equipa de peritos do FMI. | 2 |
| A negociação bolsista do outro lado do Atlântico esteve a ser penalizada pela indefinição política em Itália, que adquire novos contornos a cada dia que passa. | -2 |
| O dinheiro público aplicado na educação é um investimento e não um gasto, pois ajuda a construir um futuro mais digno para as pessoas e para o país. | 3 |
| O país sente-se manietado e não é só por isso, mas também pelo crescimento anémico, pelo investimento reduzido e por um tecido empresarial, no geral, ainda pouco competitivo. | -3 |
| Que torne mais aberto o mercado de trabalho sem cair no extremo da flexibilização dos despedimentos. | 2 |
| No meio do rebuliço, António Domingues foi o elo mais fraco e saiu, ainda que pelo seu próprio pé. | |
| Com o país transalpino cada vez mais perto de novas eleições, os investidores têm fugido dos activos de risco, provocando quedas expressivas nas acções. | -2 |
| Num comunicado emitido depois do relatório do FMI, o Ministério das Finanças" reafirma o seu empenho em prosseguir um esforço reformista, colmatando falhas passadas e projetando o futuro, consciente dos riscos que o FMI identifica como maioritariamente externos. | 2 |
| Previsão aponta para um valor semelhante ao que aconteceu em 2016, tanto em valor absoluto como em percentagem da despesa total da administração central. | 0 |
| A introdução de nova rigidez, ou a reintrodução de antigas formas de rigidez, comprometeria a competitividade e a produtividade, e dificultaria a capacidade das empresas para reagir a flutuações de procura", lê-se no comunicado. | -3 |

Tabela A.3: Extrato da dataset de frases da área económica

# Glossário

| | |
|---|---|
| Corpora | Plural de corpus, este é um conjunto de textos relacionados a um determinado assunto. |
| Dataset | Coleção de informações relacionadas e compostas de elementos separados, mas que podem ser manipulados como uma unidade pelo computador. |
| IMBd | Portal desenvolvido espeficamente para fornecer informações sobre filmes e não só. Esta plataforma fornece também dataset de opiniões dos utilizadores sobre os filmes lançados no mercados. Vários investigadores em AS já realizaram as suas experiências na base desses dados. |
| Weka | Nome inspirado a uma ave nova zelandês que não voa, só anda. Weka é uma coleção de algoritmos de aprendizagem automática para tarefas de Text Mining. Este software contém ferramentas para pré-processamento de dados, classificação, regressão, agrupamento, regras de associação e visualização. Também permite o desenvolvimento de novas técnicas de aprendizagem automática. |
| WordNet | Desenvolvido na universidade de Princeton (EUA) por George Miler, WordNet é um sistema de referência lexical on-line cujo design é inspirado nas teorias psicolinguísticas atuais da memória lexical humana. Este recurso contém cerca de 144 000 palavras inglesas organizadas em conjuntos de sinônimos, cada um representando um conceito léxico subjacente. |