Análise de sentimento em artigos de opinião

Fátima Silva
mhenri@letras.up.pt
Faculdade de Letras /Centro de Linguística da Universidade do Porto

Purificação Silvano
msilvano@letras.up.pt
Faculdade de Letras /Centro de Linguística da Universidade do Porto

António Leal
jleal@letras.up.pt
Faculdade de Letras /Centro de Linguística da Universidade do Porto

Fátima Oliveira
moliv@letras.up.pt
Faculdade de Letras /Centro de Linguística da Universidade do Porto

Pavel Brazdil
pbrazdil@inescporto.pt
Faculdade de Economia da Universidade do Porto/LIAAD – INESC Tec, Porto

João Cordeiro
jpaulo@di.ubi.pt
Universidade da Beira Interior/LIAAD – INESC Tec, Porto

Débora Oliveira
livino.debora.uporto@gmail.com
Faculdade de Economia da Universidade do Porto/LIAAD – INESC Tec, Porto

ABSTRACT: The present study, which is developed in the interface between linguistics and computer science within the framework of sentiment analysis, aims at making a computational analysis of opinion articles in the area of economics and finance. The main objectives of the study are: i) to determine the semantic orientation of text segments that express opinion by annotating the polarity (positive or negative) and the strength (scale from -3 to 3) of nouns and adjectives, and ii) to verify if a specific lexicon for the area of economics and finance has advantages in automatic annotation of sentiment over a general lexicon. To achieve these objectives, a corpus of 45 texts was selected and analyzed in 2 phases, by annotators with different training. First, a sample of 10 texts was annotated by linguists, co-authors of this paper, with the objective of developing a linguistic annotation model to ascertain the polarity and strength of words in opinion articles and extract the relevant words for this area of study. Then, a set of 35 texts was annotated by university students, replicating the annotation model developed during the first phase. Based on the linguistic annotation, the computer science team tried to establish to what extent a general sentiment lexicon for Portuguese - SentiLex - was sufficient to extract the sentiment of a sentence in a satisfactory manner or whether EconoLex, a specific sentiment lexicon, would be more efficient. The specific lexicon includes terms and multiword expressions that are relevant to the area of economics and finance and to Portuguese language, and it was developed by the authors of this study. The data was analyzed according to a blending methodology, qualitative and quantitative. The results of the analysis allow us to consider the following items as contributes of this study: i) the development of a linguistic annotation model for the analysis of the polarity and strength of the lexicon, especially of nouns and adjectives; ii) the key role, though not exclusive, of the adjectives to determine the polarity of opinion segments of the corpus articles; iii) the creation of a new specific sentiment lexicon for Portuguese in the area of economics and finance; iv) the improvement of the computational performance of EconoLex⊕SentiLex in relation to SentiLex regarding the performance in automatic annotation of sentiment. In spite of these positive results, there are some limitations, which we intend to overcome in the continuity of this interdisciplinary work, namely a more detailed linguistic analysis of the word classes that we studied, the consideration of other elements/ linguistic structures that are essential to ascertain the sentiment in NP/sentence, the extension of the corpus, the expansion of the specific lexicon of the area of economics and finance and the improvement of automatic methods for identifying evaluative words in texts of opinion and for assigning them polarity and strength.

KEYWORDS: sentiment analysis, opinion article, automatic assignment of sentiment, economics and finance, polarity and strength, lexicon, Econolex

RESUMO: O estudo apresentado realiza-se na interface entre a linguística e as ciências da computação, tendo como objetivo fazer a análise computacional de artigos de opinião na área da economia e finanças, seguindo o quadro teórico da análise de sentimento. Os principais objetivos do trabalho são i) determinar a orientação do sentimento, positivo ou negativo, e a intensidade dessa orientação através da anotação da polaridade do léxico, com incidência nos nomes e adjetivos, nos segmentos em que ocorre a expressão da opinião, e ii) verificar se um léxico específico para a área de economia e finanças tem vantagens na atribuição automática de sentimento sobre um léxico geral. Para atingir esses objetivos, foi

selecionado um corpus de 45 textos, analisado em duas fases por anotadores com formação distinta. Primeiro, uma amostra de 10 textos foi obtida e anotada pelos investigadores da área de linguística, coautores deste artigo, com o objetivo de desenvolver um modelo linguístico para determinar a orientação e intensidade da polaridade de termos em artigos de opinião e extrair termos de léxico relevantes para esta área de estudo. Em seguida, um conjunto de 35 textos foi anotado por estudantes universitários, seguindo o método utilizado na primeira amostra. Com base na anotação linguística, a equipa das ciências da computação procurou determinar até que ponto um léxico de sentimento geral para a língua portuguesa – SentiLex - é suficiente para caracterizar o sentimento de uma frase de maneira satisfatória ou se o EconoLex, um léxico específico de sentimento, seria mais eficaz. O léxico específico inclui termos e expressões multipalavra relevantes para o domínio da economia e finanças e para a língua portuguesa, e foi elaborado pelos autores deste estudo. Os dados foram analisados usando uma metodologia mista, qualitativa e quantitativa. Os resultados obtidos permitem-nos considerar os seguintes itens como contributos desta investigação: i) a elaboração do modelo de anotação linguística adotado para a análise da orientação e da intensidade da polaridade do léxico, em especial dos nomes e adjetivos; ii) o papel central, ainda que não exclusivo, dos adjetivos para a determinação da polaridade do sentimento nos segmentos opinativos dos artigos do corpus; iii) o desenvolvimento de um novo léxico de sentimento específico português para a área da economia e finanças; iv) a melhoria do desempenho computacional do EconoLex⊕SentiLex em relação ao SentiLex no que se refere ao desempenho na caracterização automática de sentimento. Apesar destes resultados positivos, há algumas limitações que constituem os elementos a desenvolver na continuidade deste trabalho interdisciplinar, nomeadamente a análise linguística mais detalhada das classes gramaticais estudadas, a consideração de outros elementos/estruturas linguísticas determinantes para a caracterização do sentimento em SN/frase, o alargamento do corpus, o aumento do léxico específico do domínio e a afinação dos métodos automáticos de identificação de termos de sentimento em textos de opinião e determinação da sua intensidade.

PALAVRAS-CHAVE: análise do sentimento, artigo de opinião, atribuição automática do sentimento, economia e finanças, polaridade e intensidade do sentimento, léxico, EconoLex

1. Introdução[1]

Este trabalho centra-se na análise computacional de artigos de opinião na área de economia e finanças, situando-se no quadro da investigação sobre análise de sentimento.

Trata-se de uma investigação realizada na interface entre a linguística e as ciências da computação, sendo objetivos centrais: i) determinar a

[1] Agradecemos aos avaliadores deste artigo os valiosos comentários, muitos dos quais foram integrados no texto. No entanto, a responsabilidade por qualquer falha remanescente é dos autores do artigo.

importância do léxico, em especial dos adjetivos e nomes, para transmitir uma opinião, positiva ou negativa, sobre o(s) tópico(s) principal(ais) do texto e ii) verificar se um léxico específico para a área de economia e finanças tem vantagens na atribuição automática de sentimento.

Através da análise do corpus, que consiste essencialmente em determinar a orientação de sentimento associado ao léxico, em especial aos adjetivos e aos nomes, e a classificação da intensidade dessa orientação[2], tanto em termos individuais como sintagmáticos, extraindo o léxico relevante do domínio temático em estudo, estabelecemos um modelo de anotação linguística para anotar um conjunto mais vasto de textos, tanto do ponto de vista manual como computacional, assim como para a constituição de um léxico específico do domínio, o Econolex.

O EconoLex é um léxico de sentimento com termos e expressões multipalavra relevantes para o domínio da economia e finanças e para a língua portuguesa, tendo sido elaborado sob a direção de alguns dos autores deste estudo, investigadores da área da linguística. A primeira versão do léxico é constituída pelos termos resultantes da anotação linguística de 10 artigos, que serviu de base para a anotação de mais 35 textos, realizada por estudantes universitários, tendo, por conseguinte, o estudo empírico relativo à previsão computacional dos valores de sentimento recaído sobre 45 textos.

Com base nos resultados da análise linguística, a análise computacional procura determinar até que ponto um léxico de sentimento geral para a língua portuguesa - SentiLex (cf., e.o., Carvalho & Silva, 2015) é suficiente para caracterizar o sentimento de uma frase de maneira satisfatória ou se o EconoLex, um léxico específico para o domínio em causa (economia e finanças), é mais eficaz.

Assim, os principais contributos desta investigação são: (i) o desenvolvimento de um modelo de anotação linguística para delimitação da

[2] Neste trabalho, seguimos a terminologia adotada por Liu (2012, 2015), que considera, como parâmetros de análise, a 'orientação do sentimento' ('sentiment orientation), positiva, negativa ou neutra, e a 'intensidade do sentimento' ('sentiment intensity'), o nível de intensidade ou força do sentimento, estreitamente relacionado com o que designa de 'classificação do sentimento' (sentiment rating'), correspondente a uma classificação discreta para exprimir a intensidade do sentimento (no caso do nosso trabalho, os níveis 3 a -3). Note-se, no entanto, que outros autores usam uma terminologia diferente e com aplicação pelo menos ligeiramente diferenciada em contexto similar, facto de que Liu dá também conta: "Sentiment orientation is also called polarity, semantic orientation, or valence in the research literature." (Liu, 2015: 21). 'Orientação semântica' é o termo usado, por exemplo, entre outros, por Taboada et al. (2011: 267-268) para referir "the polarity and strength of words, phrases, or texts".

orientação e da intensidade da polaridade do léxico em artigos de opinião; (ii) o desenvolvimento de um novo léxico (EconoLex) para Português, (iii) a comparação com o Sentilex, e (iv) o desenvolvimento de uma metodologia computacional com vista à classificação automática do sentimento de frases.

O artigo está organizado em 4 secções, a primeira das quais é a introdução. Na secção 2, fazemos um breve enquadramento teórico sobre a análise de sentimento e o género artigo de opinião. Na secção 3, descrevemos o estudo realizado caracterizando o corpus e a metodologia seguida na análise linguística e na análise computacional. A secção 4 apresenta os resultados das duas análises. Terminamos com algumas considerações finais, que nos permitem retomar o percurso efetuado e avaliar do grau de consecução dos objetivos apresentados na introdução.

2.Algumas considerações teóricas

2.1 Sobre análise de sentimento

A análise de sentimento (*sentiment analysis*) é um campo de investigação relativamente recente (Das & Chen, 2001; Tong, 2001; Turney, 2002; Pang, Lee & Vaithyanathan, 2002; Dave, Lawrence & Pennock, 2003; Nasukawa & Yi, 2003; Pang & Lee, 2008), mas que tem conhecido grande desenvolvimento (Benamara, Taboada & Mathieu, 2017). Tem como objeto de estudo as opiniões, avaliações, atitudes e emoções relativamente a certas entidades, como produtos, serviços, organizações, pessoas, eventos ou tópicos (Cambria & Hussain, 2015; Liu, 2012, 2015). Embora o número de estudos neste domínio seja muito elevado, a maior parte do trabalho na área teve como objeto a língua inglesa, havendo muito menos estudos para o Português, em especial para o Português Europeu (e.g. Carvalho, Sarmento, Silva & Oliveira; 2009; Silva, Carvalho, Costa & Sarmento, 2010; Silva & Team, 2011; Silva, Carvalho & Sarmento, 2012; Antunes, 2015; Marques Lucena et al., 2015; Forte & Brazdil, 2016).

De um modo geral, o foco principal da análise de sentimento consiste na capacidade de um sistema automático poder aferir a subjetividade e a manifestação de emoções num texto escrito em relação a certas entidades nomeadas de forma explícita ou implícita. Assim, consideramos o termo

sentimento de forma abrangente, incluindo os *afetos* e as *emoções* que a Psicologia define e caracteriza de forma rigorosa (Fiorin, 2007).

Têm sido estudadas diferentes formas de modelar as emoções, que vão desde a simples deteção de polaridade (positiva *vs*. negativa) (Turney, 2002), à combinação desta com níveis de subjetividade-objetividade (Baccianella, Esuli & Sebastiani, 2010) e até modelos mais ricos que consideram um conjunto de "emoções base" e depois definem um estado emocional como uma combinação destas emoções (Russell, 1980; Ekman, 1999).

Em termos de análise e abordagem ao problema da análise de sentimento, vários modelos e metodologias têm sido experimentados (Ravi & Ravi, 2015; Liu, 2015), quase todos tomando como base de trabalho um corpus de textos de opinião, parcialmente anotado por humanos, quanto às expressões e intensidades emocionais contidas naqueles. Estes corpora anotados permitem que sistemas de aprendizagem automática (Kodratoff & Michalski, 2014; Witten & Frank, 2016) consigam induzir conceitos gerais de identificação e caracterização de emoções em texto, ou sistemas baseados em léxico de sentimentos, que englobam os termos gerais facilitadores da identificação das emoções (e.g. Hung & Lin, 2013; Forte & Brazdil, 2016) ou combinações de ambos.

Neste contexto, a Linguística tem tido um papel crucial e várias têm sido as propostas teóricas entre as quais se destacam (cf. segundo Taboada, 2016; Benamara, Taboada & Mathieu, 2017) a teoria da avaliação (Martin & White, 2005), postura (Biber & Finegan, 1989), avaliação (Hunston & Thompson, 2000) e contrafactualidade ('nonveridicality') (Taboada & Trnavac, 2013). Apesar de não haver nenhum estudo que integre todas as componentes avaliativas disponíveis na língua para a expressão do sentimento nos textos (Benamara, Taboada & Mathieu, 2017), em grande parte pela sua complexidade e abrangência, é possível identificar alguns dos elementos mais relevantes na análise de sentimento na investigação em curso.

Um desses elementos, e um dos que tem recebido maior enfoque, é o léxico, e, em particular, as classes dos adjetivos, dos nomes e, em menor número, os verbos (e.g. Levin, 1993; Mathieu, 2005; Taboada, Anthony & Voll, 2006; Freitas, 2013). Estas classes de palavras têm sido analisadas quanto a polaridade positiva, negativa e neutra, muitas vezes associada a

escalas com diferentes valores, que variam de acordo com as propostas (Goldberg & Zhu, 2006; Pang & Lee, 2008). Uma das tarefas mais complexas relacionada com a atribuição de valores de polaridade a adjetivos, nomes e verbos de sentimento e com a criação de um léxico com estes valores (Levin, 1993; Neviarouskaya, Prendiger & Ishizuka, 2009; Freitas, 2013) surge sempre que o significado base dos itens não é evidente e/ou depende do contexto em que é usado.

2.2. Sobre o género artigo de opinião

O artigo de opinião é um género de discurso que se caracteriza pelo seu traço marcado de comentário (Adam 1997; Charaudeau 2006; Cunha 2012). No domínio jornalístico, consiste essencialmente na discussão, por parte de um autor, mais ou menos especializado, de um assunto atual e considerado de relevância para o leitor, ocupando um espaço que oscila habitualmente entre meia página e uma página no jornal, numa secção dedicada a questões de economia e finanças, no caso dos textos que analisámos. Este autor, é, de acordo com Cunha (2012: 75), "um especialista externo à instância mediática comentando um facto ou mesmo provocando a sua emergência". Nesse sentido, existe uma assimetria entre a instância que produz o texto e a instância que o recebe, na medida em que a primeira é institucionalmente legitimada na exposição do seu ponto de vista enquanto a segunda o recebe de modo a informar-se e esclarecer-se sobre o tópico em questão (Rodrigues, 2005; Cunha, 2012). Tipicamente, o objeto de análise é um tema que se centra sobre factos recentes, sobre os quais é frequente a geração de diferentes pontos de vista, sendo, por conseguinte, polémico. Assim, consiste frequentemente num facto que "mobilizou a atenção da opinião pública e agora pede que os veículos de comunicação apresentem as análises "esclarecidas" de especialistas da área em que o fato se deu" (Cunha, 2012: 76).

Tendo em consideração que a simples explicitação da opinião do autor do texto não é suficiente para garantir a adesão do leitor ao seu ponto de vista, até pelo facto de existirem habitualmente pontos divergentes sobre o tema, são desenvolvidas estratégias de argumentação que visam convencer o leitor deste ponto de vista, fazendo-o aderir à leitura proposta sobre o

tema objeto de discussão.

A textualização das estratégias argumentativas no artigo de opinião, a partir da qual se pode analisar a orientação do sentimento ou da opinião, em termos da sua polaridade positiva, negativa ou neutra, faz uso de vários recursos linguísticos, entre os quais o léxico. A análise decorrente da descrição do léxico fornece pistas para a determinação da polaridade do documento no seu todo, como expressando sentimento positivo ou negativo (cf. Silva *et al*. 2015, 2018). É verdade que a sua combinação com outros aspetos da análise de sentimento deve ser tida em conta, porque a expressão de sentimento no texto recorre a outros elementos de natureza linguística e discursiva, que não se esgotam nas categorias gramaticais consideradas. No entanto, é indiscutível que a análise semântica lexical e frásica dos nomes, adjetivos e verbos e da sua combinatória em sintagmas fornece indicações relevantes para a análise da expressão do sentimento em textos de opinião do domínio em análise.

3. O estudo

3.1. Corpus

O estudo tem como base um corpus de 45 textos. Numa primeira fase, foi analisada uma amostra de 10 textos do domínio discursivo da economia e finanças do género artigo de opinião, para teste da metodologia de análise linguística e extração dos termos a integrar no EconoLex, o léxico específico para a área de economia e finanças. Esses textos foram extraídos de forma aleatória de três jornais digitais diários, *Público* (1), *Diário de Notícias* (1) e *Jornal de Negócios* (8). Enquanto os dois primeiros jornais referidos são generalistas, o terceiro é um jornal especializado em assuntos de economia e finanças. A recolha foi realizada entre 3 de janeiro e 8 de fevereiro de 2017, tendo sido subordinada a uma unidade temática, a Caixa Geral de Depósitos, para evitar uma dispersão do léxico e aferir os valores do léxico de sentimento encontrado em contextos similares.

Numa segunda fase, foram considerados mais 35 textos aos quais foi aplicada a mesma metodologia de análise da amostra. Estes textos foram igualmente selecionados de forma aleatória, a partir de jornais generalistas e

diários - *Público* (3), *Jornal de Notícias* (2), *Diário de Notícias* (2) e *Observador* (1) -; de jornais especializados, diário e semanário, respetivamente - *Jornal de Negócios* (1) e *Jornal Económico* (12) -; de um jornal generalista semanário - *Expresso* (9) -; e de uma revista semanal - *Visão* (5). Ao contrário da amostra, a recolha destes textos não seguiu um critério temático, na medida em que se pretendia alargar o léxico de sentimento, embora se mantivesse sempre no domínio de economia e finanças.

3.2. Metodologia de Análise Linguística

Feita a seleção e extração do corpus, a análise de sentimento dos 10 artigos de opinião referidos em 3.1. realizou-se em várias etapas:

1. Extração dos segmentos dos textos em que se exprime opinião, isto é, em que está implicado um sentimento positivo ou negativo;
2. Delimitação, nos segmentos considerados, das frases relevantes para análise, isto é, dos contextos em que ocorre uma expressão de sentimento;
3. Determinação das categorias que exprimem sentimento – nomes, adjetivos e multipalavras – e dos sintagmas nominais por eles constituídos;
4. Anotação da orientação do sentimento de cada uma destas categorias – positiva, negativa ou neutra;
5. Anotação da intensidade da polaridade dessa orientação, numa escala de 3 a -3 valores;
6. Avaliação dos resultados da análise;
7. Identificação das palavras relevantes para integração num léxico específico do domínio da economia e finanças, o EconoLex.

A anotação do corpus foi feita manualmente por 3 investigadores especialistas em linguística. Nos casos em que a anotação deu resultados discrepantes em termos da orientação de sentimento e, sobretudo, da intensidade da polaridade dessa orientação, foi realizada uma reanálise e reavaliação conjunta dos dados em questão para aferição de um valor. As palavras e os sintagmas anotados foram sujeitos a inquérito junto de dois informantes especializados em linguística, mas externos à investigação em curso, para aferição da anotação efetuada pelos investigadores, em

geral, e discussão dos casos de anotação não concordante, em particular[3]. A anotação da orientação e da intensidade de sentimento baseou-se nos seguintes critérios: (i) avaliação por falantes nativos, especialistas e não especialistas; (ii) consulta do significado lexical dos termos em dicionários gerais; (iii) e consulta de dicionários gerais de sentimento de Português e de outras línguas (Sentilex, So-Cal, entre outros).

3.2.1 Identificação dos segmentos relevantes

Como referimos, a análise dos textos constitutivos do corpus consistiu, em primeiro lugar, na identificação dos segmentos em que se veicula de forma explícita (cf. (1)) ou implícita (cf. (2)) a expressão de opinião sobre o tema relevante, tendo sido excluídos os segmentos em que a expressão de sentimento não se centra sobre este tópico de modo direto (cf. (3)). Assim, em (1), a expressão de opinião é marcada, nomeadamente, através do verbo de opinião 'julgar' e do SN 'infeliz marca'. Por sua vez, em (2), a opinião é dada através de uma analogia com um filme descrito no 1.º parágrafo do texto. Finalmente, em (3), o segmento de opinião não se refere diretamente à Caixa Geral de Depósitos, mas às declarações das duas entidades citadas.

(1) Tem-se falado de prémios de gestão. Prémios de gestão avultados são uma infeliz marca do sistema financeiro. Não julgo apropriada, e até prejudicial, a existência para a CGD de prémios da natureza dos da banca privada, em função dos lucros. Os prémios estão na origem dos males do sistema financeiro, colocando na mira dos gestores bancários a perspetiva de ganhos pessoais elevadíssimos na dependência dos lucros das instituições. [texto 8]

(2) É um bonito dia de sol que nos acolhe no início do filme "La La Land". Estamos em Los Angeles e o trânsito está engarrafado numa auto-estrada e nada se move. Vamos seguindo então a música diferente que sai de cada carro parado. É uma sinfonia pouco sincronizada, porque cada um escuta uma coisa diferente. Quando paramos, a

3 Agradecemos a Idalina Ferreira e a Luís Filipe Cunha a colaboração na anotação da polaridade do léxico extraído dos 10 textos analisados na primeira fase do trabalho.

> música fica mais alta. Então a personagem sai do carro e começa a dançar. Rapidamente todos os outros condutores saltam para as capotas dos carros e dançam sem parar. Todos dançam como se fossem Gene Kelly em "Singing in the Rain". Os musicais são assim. Serão sempre. A Caixa Geral de Depósitos é uma "La La Land" à nossa dimensão... [texto 4]

(3) As declarações de Fernando Faria de Oliveira e Carlos Santos Ferreira na comissão parlamentar de inquérito à Caixa Geral de Depósitos (CGD) vieram enfatizar aquilo que há muito é uma evidência: o banco só será pacificado quando se realizar e forem tornadas públicas as conclusões da auditoria forense, aprovada pela Assembleia da República a 20 de Julho de 2016, mas que ainda se encontra numa qualquer gaveta do Banco de Portugal. [texto 5]

3.2.2 Identificação de frases e categorias de palavras a analisar

No âmbito dos segmentos selecionados, foram identificadas as frases relevantes para a análise de sentimento, nas quais estavam contidas as categorias sintáticas analisadas: nomes e adjetivos. Neste contexto, foram considerados não só nomes e adjetivos que exprimem intrinsecamente sentimento (cf. (4) e (5)), mas também nomes e adjetivos que estão relacionados com o domínio específico da economia e finanças e são usados na textualização de sentimento, independentemente de terem, no seu significado lexical nuclear, uma polaridade neutra (cf. (6) e (7)).

(4) Tal como não é novo que a Caixa vai cortar o número de balcões e de funcionários de forma **agressiva**: são menos 2200 trabalhadores e perto de 200 balcões.

(5) Os prémios estão na origem dos **males** do sistema financeiro, colocando na mira dos gestores bancários a perspetiva de **ganhos** pessoais elevadíssimos na dependência dos lucros das instituições.

(6) A **gestão** é pouco qualificada e a mão-de-obra também.

(7) A escolha de Macedo foi inteligente, tanto do ponto de vista **técnico** como **político**.

Assim, enquanto o adjetivo 'agressiva' e os nomes 'males' e 'ganhos' exprimem no seu significado lexical nuclear a expressão de um sentimento, o nome 'gestão' e os adjetivos 'técnico' e 'político' não contêm esse traço no seu significado lexical básico, sendo necessário atender ao significado lexical atribuído contextualmente para legitimar a sua consideração na delimitação de valores de sentimento a uma dada expressão ou frase. Isto não significa, no entanto, desvalorizar o significado contextualmente atribuído a nomes e adjetivos tipicamente ocorrentes em léxicos de sentimento, dado que a orientação de sentimento e da intensidade dessa orientação é determinada na articulação do significado lexical nuclear com o significado contextualmente adquirido, em particular quando se considera o sintagma nominal (SN) em que ocorrem e o contexto frásico integral.

No que se refere aos adjetivos, foram considerados os seguintes critérios de análise: função - predicativa ou atributiva - e, em relação a esta última, foi definida a posição pré-nominal ou pós-nominal. Os exemplos (8) e (9) atestam a diferença de posição do adjetivo (atributiva e predicativa), enquanto os exemplos (10) – (11) exemplificam a posição pré ou pós-nominal dos adjetivos em posição atributiva.

(8) A dívida ingerível, o crescimento **anémico**, os impostos **altos**, o investimento **nulo**, a regulação **ineficaz**, a separação dos portugueses entre protegidos e excluídos exigem mais do que um simples "virar de página" da austeridade. [texto 2]

(9) A escolha de Macedo foi **inteligente**... [texto 3]

(10) Esse é, aliás, o **grande** alerta da OCDE ao referir que o problema da banca está longe de estar resolvido. [texto 1]

(11) São necessárias taxas de crescimento de 3%, 4% ou 5%, ou seja, taxas **ambiciosas**... [texto 1]

Por outro lado, foram analisadas as subclasses dos adjetivos – *adverbiais, numerais, qualificativos* e *relacionais*[4]. Os exemplos (12)

[4] Na análise das classes semânticas, seguimos a proposta de Ferreira (2013). Agradecemos à autora a colaboração na anotação semântica dos adjetivos relevantes para a análise de sentimento proposta.

– (15) ilustram, respetivamente, cada uma das subclasses de adjetivos atualizadas nos adjetivos 'fortes', 'última', 'ágil' e 'agressiva', 'estruturais'. Os *adverbiais* são genericamente aqueles que podem ser transformados em advérbios. No exemplo (12) isso acontece por causa da posição do adjetivo. Se o adjetivo estivesse em posição pós-nominal seria *qualificativo* ("recapitalizações fortes"). Os *numerais* relacionam-se com ordenação (cf. (13)); os *qualificativos* atribuem propriedades e são graduáveis (cf. (14)); os *relacionais* derivam em geral de nomes ('estrutura' - 'estrutural') (cf. 15)). Note-se que, se usarmos os conceitos da área de análise de sentimento, o adjetivo *adverbial* no exemplo (12) representa um *intensificador*.

(12) …exigiram aos seus bancos que procedessem a **fortes** recapitalizações dos seus capitais. [texto 9]

(13) Esta é a **última** oportunidade para a Caixa. [texto 3]

(14) Na prática, referia-se ao atraso com que o banco aumentou 'spreads' e comissões e cortou remuneração dos depósitos, face à concorrência, muito mais **ágil** e **agressiva**. [texto 5]

(15) Está na altura de Portugal e a UE se capacitarem que os problemas **estruturais** se resolvem com medidas de fundo. [texto 9]

3.2.3. Expressões Multipalavra

Além de nomes e de adjetivos, foi também considerada a ocorrência de multipalavras, de que se dá exemplo nos segmentos (16) e (17), com as multipalavras 'banco público' e 'sistema político'.

(16) Na verdade, a vantagem para a economia que se pode tirar da existência de um **banco público** é ser diferente na sua política de aplicações. [texto 8]

(17) … o défice está em dia, o diabo não veio, não há muitas greves e experimentamos a genialidade de um **sistema político** "plástico" que impediu o vazio do poder. [texto 2]

Entende-se que uma multipalavra é "uma sequência de palavras que

atuam como uma simples unidade em algum nível da análise linguística" (Calzolari *et al.*, 2002: 1934), pode englobar um variado número de construções, tais como expressões fixas, compostos nominais e construções verbo-partícula (Sag *et al.*, 2002: 191) e ocorre com frequência em domínios técnicos. O conceito de *multipalavra* é complexo tanto do ponto de vista da análise linguística como computacional. Neste trabalho, de natureza exploratória, não foi nosso objetivo rever de forma aprofundada este conceito, mas usamo-lo no seguimento de McCarthy (1990); Nattinger & DeCarrico (2001); Calzolari *et al*. (2002); Sag *et al*. (2002); Ranchhod (2003); Gómez Molina (2004); Thornbury (2007); Abalada et al. (2010); entre outros.

No corpus, a delimitação de multipalavras da área de economia e finanças fez-se considerando sobretudo a existência de expressões lexicalizadas. Verificamos que, sempre que ocorre uma multipalavra constituída pela estrutura 'nome + adjetivo', o adjetivo é relacional, como é o caso da sua ocorrência nos exemplos (16) e (17), respetivamente 'banco público' e 'sistema político'.

3.2.4. Anotação do Corpus

Com base nestas categorias (nomes, adjetivos e multipalavras), procedemos à análise do corpus, cuja anotação se fez numa grelha considerando oito parâmetros: contexto, frase relevante, polaridade do nome, polaridade do adjetivo, posição do adjetivo, subclasse semântica do adjetivo, polaridade do SN e polaridade da frase. A tabela 1 contém a exemplificação da grelha usada para a anotação.

A tabela 2 mostra apenas exemplos de algumas destas frases com os respetivos valores de sentimento atribuídos (colunas um e dois da tabela). Os valores de sentimento variam de -3 (fortemente negativo) a 3 (fortemente positivo), isto é, podem incluir qualquer valor do conjunto {-3, -2, -1, 0, 1, 2, 3}. Nas colunas três e quatro da tabela podemos ver os valores de sentimento associados aos termos uni-palavra que ocorrem nas frases, divididos pelas classes gramaticais de 'nome' e 'adjetivo' (ex. 'crise', 'anémico'). As expressões multipalavra são maioritariamente do tipo 'nome + adjetivo', como, por exemplo: 'carga fiscal', 'investimento público',

'sistema político'.

Note-se que, na determinação da intensidade da polaridade, é necessário ter em conta que, além do conhecimento linguístico e da competência textual dos potenciais anotadores e, de forma mais global, dos leitores, intervém o que poderíamos designar por conhecimento do mundo e um sistema de crenças de natureza diversa, na base do que poderíamos discutir em termos da distinção entre uma polaridade mais subjetiva, dependente da ideologia ou da conceção do mundo do falante/ouvinte, e uma outra, mais universal/objetiva. Este sistema de crenças, socioculturalmente marcado, pode interferir na polaridade atribuída (cf., e.o., Li & Liu 2012: 128). A título de exemplo, consideremos o SN "pensamento conservador", que ocorre na tabela 2. Embora o valor negativo seja claro no texto, é preciso ler o parágrafo e interpretar a ideologia económica do autor para lhe atribuir esse valor negativo. Noutro contexto ideológico, de índole mais conservadora e neoliberal, essa mesma frase nominal poderia eventualmente ter valor positivo.

As questões colocadas por este tipo de diferenças podem ser limitadas por vários procedimentos como sejam a anotação por um conjunto alargado de informantes e o recurso a léxicos gerais e específicos; a consideração de outros elementos que concorrem para a configuração do sentido dos segmentos de opinião; a análise da percentagem de concordâncias obtidas entre anotadores, do ponto de vista linguístico, e a aplicação de algoritmos computacionais. Estes foram os critérios que nortearam a nossa anotação do corpus já referido.

TABELA 1: Grelha de anotação manual da análise linguística de sentimento

contexto	**frase relevante**	**nome: polaridade**	**adjetivo: polaridade**	**posição predicativa/atributiva**	**posição atributiva: pré/pós-nominal**	**adjetivo: subclasse**	**SN: polaridade**	**polaridade da frase**
Portugal não pode estar a governar só para os mercados, ou seja, para tentar demonstrar que o défice está melhor. Mas a verdade também é que, se não baixar o défice, é penalizado na dívida.	Portugal não pode estar a governar só para os mercados, ou seja, para tentar demonstrar que o défice está melhor.	*défice* -1	*melhor* +2	pred.	-	qualif.	pos.	neg.
Mas a verdade também é que, se não baixar o défice, é penalizado na dívida. 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.	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.	*país* 0	*manietado* (particípio) -2	pred.	-	qualif.	neg.	neg.
		crescimento +2	*anémico* -2	atrib.	pos	qualif.	neg.	
		investimento +1	*reduzido* -1	atrib.	pos	qualif.	neg.	
		tecido 0	*empresarial* 0	atrib.	pos	relac.	neutro	
		tecido 0	*competitivo* +1	atrib.	pos	qualif.	neg.	

TABELA 2 - Exemplos de frases, termos relevantes e valores de sentimento

Frase (F)	V(F)	V(nome)	V(adjetivo)	V (express)
Face às eleições anteriores, o Bloco perdeu metade dos votos, metade dos deputados e entrou numa **crise** profunda.	-3	*crise*: -2	*profunda*: -2	*crise profunda*: -3
Por mais avanços que a tecnocracia europeia se mostre disposta a dar, na sequência da crise do euro, a predominância de um *pensamento conservador* nas esferas de poder torna previsível que fossem impostas de novo *medidas violentas* de **contenção** *orçamental*.	-2	*pensamento*: 0 *medidas*: 0 *contenção*: -1	*conservador*: 0 *violentas*: -1 *orçamental*: 0	*pensamento conservador*: -1 *medidas violentas*: -2 *contenção orçamental*: -1
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	*manietado*: -2 *crescimento*: 2 *investimento*: 1	*anémico*: -2 *reduzido*: -1 *competitivo*: 1	*crescimento anémico*: -1 *investimento reduzido*: -1

A anotação e análise dos 10 artigos iniciais deram origem à primeira versão do EconoLex. Uma outra parte do léxico, bastante significativa, resultou da anotação de mais 35 artigos por estudantes do ensino superior. Assim, no total, foram usados 45 artigos da área da economia e finanças e, após uma pré-análise, foram escolhidas 370 frases.

3.3. Análise computacional

Para este estudo, usamos uma metodologia que segue uma abordagem baseada em métodos computacionais, através do uso de corpora e léxicos eletrónicos, bem como procedimentos que envolvem cálculo de valor previsto de sentimento de frases.

3.3.1. Uso de léxico no cálculo de valor de sentimento

Foi implementado um procedimento automático para caracterizar o sentimento de uma determinada frase. Para cada palavra da frase, é feita uma procura no léxico. Caso a palavra ocorra, tomar-se-á o valor de sentimento associado à mesma. Neste procedimento, o cálculo final do valor de sentimento da frase é processado em duas fases. Na primeira, o procedimento elabora uma soma dos valores de todos os termos encontrados no léxico. Na segunda fase, estes valores são reescalonados, dividindo-os por um fator que tem um efeito equalizador, tendo em conta certas gamas de valores de sentimento, para os diferentes léxicos considerados. No final, cada frase fica associada a um valor numérico que traduz o sentimento geral dessa frase e que poderá ser negativo, neutro (zero), ou positivo, em diferentes níveis de intensidade. Este processo está descrito com mais pormenor na secção 3.3.2.

Léxico geral SentiLex

O SentiLex (Silva *et al*., 2012), mais precisamente a versão SentiLex-PT02 que foi usada neste estudo, foi desenvolvido para uso geral (i.e. é *general-purpose lexicon*). O léxico contém uma lista de cerca 82.347 entradas, que são palavras flexionadas ou expressões idiomáticas e multipalavras em Português. Este conjunto corresponde a 7.014 termos lematizados. Cada entrada está associada a um valor de sentimento do conjunto {-1, 0, 1}, significando respetivamente valores de sentimento *negativo*, *neutro* e *positivo*. Assim, por exemplo, para a entrada 'abafada' e 'combaterá o desemprego' temos respetivamente os valores -1 e 1 associados, sendo a primeira um adjetivo, cujo lema é 'abafado', e a segunda uma expressão multipalavra, que na área de lexicologia é referida como *colocação*. Na

tabela 3 encontramos algumas entradas do SentiLex, com os respetivos lemas, classe gramatical e valor de sentimento associado.

TABELA 3 - Entradas típicas no SentiLex

Termo	Lema	Classe	Valor
abafada	abafado	Adj	-1
condenou	condenar	Vrb	-1
votou	votar	Vrb	0
encorajando	encorajar	Vrb	1
irregularidade	irregularidade	Nom	-1
combaterá o desemprego	combater o desemprego	Multi	1

Ao trabalhar com o SentiLex, notou-se que este não inclui alguns termos importantes da área da economia e finanças, como, por exemplo, 'inflação', 'despedir', e 'incumprimento', entre outros. Na versão lematizada, a maioria dos termos do SentiLex são adjetivos. No entanto, quando se considera a versão flexionada, como seria previsível, há mais verbos do que adjetivos (ver tabela 4).

TABELA 4 - Distribuição de ocorrências de lemas por tipo (classe)

Classe	Lema	Forma flexionada
Nome	1,080	1,280
Adjetivo	4,779	16,863
Verbo	489	29,504
Multipalavra	666	34,700
Total	7,014	82,347

A maioria das entradas (66%) tem o valor negativo (-1), enquanto só um quarto (25%) tem o valor positivo (1) e as restantes entradas (9%) têm o valor neutro (0).

Léxico específico EconoLex

O EconoLex é, tal como foi referido anteriormente, um léxico de sentimento com termos e expressões multipalavra relevantes para o domínio da economia e finanças em Português. Este léxico contém menos entradas do que o SentiLex, pois inclui 2811 termos não flexionados, que correspondem a 1246 termos lematizados. Neste léxico os valores positivos e negativos estão mais equitativamente distribuídos, relativamente ao SentiLex, pois há 46% de termos positivos, 35% de negativos e 19% de neutros.

3.3.2. O cálculo final do valor de sentimento

O léxico específico EconoLex foi usado como extensão do léxico geral SentiLex. A junção dos dois léxicos é aqui designada de EconoLex⊕SentiLex.

Como um dado termo pode ocorrer em ambos os léxicos, torna-se necessário decidir qual o valor de sentimento a aplicar na caracterização do sentimento de uma frase. Assim, para cada termo $t \in$ {EconoLex, SentiLex} escolhe-se o *valor*(*t*) no primeiro ou no segundo léxico? Decidimos dar preferência ao léxico específico EconoLex, atendendo a que este caracteriza de forma mais adequada o domínio do texto em análise. Em termos procedimentais, esta preferência é executada dando prioridade aos termos do léxico EconoLex sobre o léxico SentiLex.

Cada léxico é ordenado de acordo com o tamanho da expressão que corresponde ao número de palavras envolvidas. As expressões de tamanho maior são colocadas no topo da lista de prioridades relativamente às expressões de tamanho inferior e aos termos uni-palavra que aparecem mais para o final da lista. Esta ordenação tem como objetivo aplicar o valor da expressão multipalavra, caso ocorra, em vez dos valores das uni-palavras que a constituem. Considere-se, por exemplo, que se pretende obter o valor do seguinte fragmento, usando o léxico na Tabela 5 e concentrando especial atenção no segmento 'crescimento anémico':

(18) 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.

TABELA 5 - Esboço de entradas no léxico

Termo	Tamanho	Valor
crescimento anémico	2	-1
anémico	1	-2
crescimento	1	2

Tendo em conta a ordem das entradas na tabela, o valor do segmento corresponde à primeira entrada: 'crescimento anémico' com o valor -1. Caso não fosse aplicada esta ordem de prioridade, aplicar-se-iam os valores dos termos que constituem esta expressão, isto é, os valores atribuídos individualmente a 'anémico' e a 'crescimento'. Repare-se que, nesse caso, o valor deste segmento seria obtido através da soma dos seus valores individuais, resultando numa neutralização (zero) do valor final do segmento, o que seria claramente inapropriado.

3.3.3. Reescalonamento de valores de sentimento

Como o objetivo é prever os valores de sentimento usando um método computacional baseado no léxico, é necessário garantir que os valores gerados são da mesma grandeza e, se não o forem, ajustá-los e integrá-los da forma mais conveniente. Para atingir este fim, usámos um reescalonamento de valores, tendo por base os desvios padrão de pares de populações, sendo cada população caracterizada pela aplicação de um determinado léxico, no cálculo do sentimento das frases.

Em primeiro lugar, usámos o nosso procedimento para gerar as classificações de sentimento das frases. Os valores gerados foram usados para construir um histograma que pode ser caracterizado pela sua *média* e *desvio de padrão* σ_{prev}. Os valores de σ_{prev} para SentiLex, EconoLex e EconoLex⊕SentiLex foram respetivamente: 1,02, 3,65 e 3,85. Repare-se no baixo valor de σ_{prev} para o SentiLex e que se deve fundamentalmente à gama de valores de sentimento mais reduzida neste léxico ({-1,0,1}). Um processo semelhante pode ser usado para caracterizar os valores atribuídos pelos especialistas humanos que classificaram manualmente o valor de

sentimento para cada frase. Assim, vamos obter um histograma semelhante, que pode ser caracterizado por σ_{true}. Neste caso, o valor obtido foi de 1,45.

Subsequentemente, decidiu-se tomar como fator de reescalonamento o rácio $\sigma_{prev}/\sigma_{true}$. No caso de EconoLex, por exemplo, o fator é de 3,65/1,45 = 2,51. Após o cálculo do fator, todos os valores previstos são divididos pelo respetivo fator, sendo transformados para uma escala normalizada.

3.3.4. Avaliação de previsões de polaridade

Em termos de avaliação da eficácia de uma abordagem, o método convencional utilizado nas áreas de *Aprendizagem Automática* (*Machine Learning*), *Prospeção de Dados* (*Data Mining*) e *Prospeção de Texto* (*Text Mining*) consiste na comparação das previsões da abordagem com os valores corretos (*golden standard*). Assim, neste estudo também foram comparados os valores de sentimento calculados para uma frase f, $ValS_{prev}(f)$, com os valores de sentimento atribuídos pelos especialistas para essa mesma frase, $ValS_{true}(f)$, que se considera como sendo o *correto*. Deste modo, calculou-se a diferença absoluta (erro absoluto) entre os dois valores, nomeadamente o erro absoluto (EA) para cada frase f:

$$EA(f) = | ValS_{prev}(f) - ValS_{true}(f) |.$$

Por exemplo, se $ValS_{prev}(f) = 1$ e $ValS_{true}(f) = -1$, então o valor de EA é igual ao valor 2 (1-(-1)). Pretende-se que a diferença dos dois valores seja a mais baixa possível. Esta medição é aplicada a todas as frases e depois é calculada a média dos erros absolutos, ou o chamado *Erro Absoluto Médio* (*EAM*), como medida de desempenho geral.

3.3.5. Metodologia de validação cruzada com 'deixa-um-fora'

Na área da *Aprendizagem Automática*, os procedimentos propostos (e.g. um classificador) são normalmente avaliados seguindo a metodologia designada de *validação cruzada*. Trata-se de um conjunto de dados que normalmente é dividido em dados de treino e de teste, sendo o primeiro usado para construir um modelo para gerar os valores previstos.

Seguidamente, mede-se o desempenho do procedimento nos dados de teste.

Uma versão mais elaborada deste processo consiste em dividir o conjunto de dados em N blocos de tamanho aproximadamente igual (normalmente N = 10) e executar N ciclos treino/teste. Para o i-ésimo ciclo, o procedimento é treinado com todos os blocos, exceto o bloco i que é usado para teste. O desempenho final é obtido a partir da média do desempenho nos N blocos.

No contexto deste trabalho, um bloco de dados é simplesmente um bloco de textos. Aqui o "treino" envolve a elaboração do léxico a partir de N-1 blocos (textos), bem como o reescalonamento descrito na subseção 3.3.3.

A metodologia *Validação Cruzada com Deixa-Um-Fora (Leave-One-Out Cross-Validation, LOOCV)* é semelhante à descrita anteriormente, mas o número de blocos N é igual ao número de casos (textos) nos dados. Neste estudo, foi esta a metodologia de avaliação adotada. Assim, usámos N-1 textos (no nosso caso 44 textos) para elaborar o léxico específico e o texto que ficou de fora foi usado para avaliar as previsões. O processo é repetido N vezes (i.e., 45 vezes) e toma-se a média dos resultados de desempenho como desempenho final, medido em termos de EAM.

4. Discussão dos resultados

4.1. Análise linguística

Tendo explicitado na secção 3.2. os parâmetros usados na análise do corpus, apresentamos nesta secção os resultados dessa análise, subdividindo-os em duas componentes. Na primeira, apresentamos a análise dos dados; na segunda, descrevemos as regularidades encontradas no corpus no que se refere à análise de sentimento e à sua orientação.

4.1.1. Dados quantitativos

O corpus dos 10 artigos de opinião perfaz no seu total 5407 palavras, o que equivale a uma média de 540,7 palavras por texto. Tendo sido contabilizadas as ocorrências de nomes (N) e de adjetivos (A) e de multipalavras (MP)

no corpus, verificamos que elas correspondem, respetivamente, a 1,96%, 2,75% e 0,61%, o que constitui uma percentagem globalmente pequena de ocorrências lexicais na totalidade do corpus, embora forneça um contributo fundamental para a argumentação veiculada na explicitação da opinião do autor. A tabela 6 especifica estes resultados, assim como a sua distribuição por texto analisado.

TABELA 6: Quantificação das ocorrências relevantes em função do número total de palavras

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	totais
n.º total de ocorrências de palavras	411	1144	425	345	496	344	455	716	585	486	5407 (100%)
n.º de ocorrências N	15	19	10	5	9	3	6	11	11	17	106 (1,96%)
n.º de ocorrências ADJ	18	21	13	7	14	3	7	19	24	23	149 (2,755%)
n.º de ocorrências de multipalavras	1	3	1	0	2	0	0	3	2	2	14 (0,61%)

Na tabela 7, é explicitada a quantificação dos adjetivos em função da sua subclasse.

TABELA 7: Quantificação dos adjetivos ocorrentes no corpus por subclasse semântica

	tipos	**n.º de ocorrências**	**%**
adjetivos 149	adverbiais	11	7,39
	numerais	1	0,67
	qualificativos	110	73,82
	relacionais	27	18,12

Neste âmbito, verificamos que há uma dominância de adjetivos qualificativos no contexto da expressão de valores de sentimento (73,82%),

logo seguida pelos adjetivos relacionais (18,12%). Com um valor menos expressivo, 7,39%, ocorrem os adjetivos adverbiais, sendo residual o número de adjetivos numerais nas ocorrências relevantes. Esta tendência exprime, de forma evidente, o papel dos adjetivos qualificativos no contexto da expressão de valores de sentimento, decorrendo das suas propriedades semânticas de expressão de uma qualidade, o que, no contexto do presente estudo, se coaduna com a manifestação da opinião do autor do texto relativamente a determinadas entidades (essencialmente denotadas pelos nomes). A ocorrência de adjetivos relacionais em número relativamente elevado está, em certa medida, ligada à sua associação a nomes com os quais, em contextos vários, produzem expressões lexicalizadas do tipo multipalavras.

Quanto ao adjetivo, verificamos que há uma diferença percentual significativa em relação à sua posição, que é sobretudo atributiva, com 81,8% das ocorrências, o que vai ao encontro da sua função central na expressão da avaliação da polaridade das entidades textuais, frequentemente denotadas por nomes da área da economia e finanças. Relativamente à posição dos adjetivos com função atributiva, observamos que 61,44% das suas ocorrências nesta posição é pós-nominal, o que constitui mais de metade das ocorrências. Estes dados podem ser observados de forma circunstanciada na tabela 8.

TABELA 8: Quantificação da ocorrência de adjetivos quanto à posição

		T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	totais
n.º de ocorrências ADJ	total	18	21	13	7	14	3	7	19	24	23	149 (100%)
	predicativa	6	1	3	1	3	0	1	6	3	3	27 (18,12%)
	atributiva pré-nominal	1	4	1	1	6	0	2	3	4	8	30 (20,134)
	atributiva pós-nominal	11	16	9	5	5	3	4	10	17	12	92 (61,744)

4.1.2. Regularidades encontradas na expressão dos valores de sentimento

De seguida, apresentamos algumas regularidades identificadas através da análise do corpus.

Nos casos em que a polaridade dos nomes e dos adjetivos é distinta entre si, é o adjetivo que é responsável pela polaridade do SN, como se verifica nos exemplos (19) e (20), em que é o adjetivo a determinar a polaridade, respetivamente, positiva e negativa do SN, quer em posição atributiva, quer em posição predicativa.

(19) Portugal não pode estar a governar só para os mercados, ou seja, para tentar demonstrar que o **défice** está **melhor**. [texto 1]
N -1 Adj 2 SN Pos

(20) Que encare **verdades inconvenientes**, como o fracasso a prazo do sistema de pensões. [texto 2]
N 3 Adj -2 SN Neg

Há casos em que não é possível determinar o contributo do adjetivo para a valoração do SN, pois ambos têm a mesma polaridade, como é o caso dos exemplos (21) e (22).

(21) A **dívida ingerível**, o crescimento anémico, os impostos altos, o investimento nulo, a regulação ineficaz, a separação dos portugueses entre protegidos e excluídos exigem mais do que um simples «virar de página» da austeridade. [texto 2]
N -3 Adj -3 SN Neg

(22) O que a OCDE contrapõe é uma dúvida a esse estado de **pura felicidade**. [texto 2]
N 3 Adj 3 SN Pos

Por outro lado, os adjetivos qualificativos têm, geralmente, valoração positiva ou negativa, e os relacionais, valoração neutra. Os adjetivos qualificativos 'excelentes' (cf. (23)) e 'chocante' (cf. (24)) são adjetivos qualificativos com especificação da orientação de sentimento, positivo e negativo, respetivamente, mas ao adjetivo 'políticos' (cf. (25)), de natureza relacional, atribui-se valoração neutra.

(23) Esta partidarização da CGD foi visível em administrações sucessivas, onde no meio de **excelentes gestores**, existiam prateleiras de luxo para quem saía da órbita governamental. [texto 4]

(24) Este **jogo de sombras** é **chocante**. [texto 5]

(25) O que passa por se proteger dos **lóbis políticos** e alterar de forma substantiva a sua actividade creditícia. [texto 3]

Apesar de, como se disse os adjetivos serem maioritariamente responsáveis pela polaridade do SN, há casos em que são subespecificados quanto à valoração em virtude de o conteúdo semântico do nome ser decisivo para atribuição da polaridade, como acontece nos exemplos (26) e (27). Neste contexto, o adjetivo 'sério' pode assumir uma polaridade positiva ou negativa conforme o nome ao qual se associa. Assim, em (26), ligado ao nome 'forma' e dado o cotexto, assume um valor positivo, sendo a intensidade da polaridade que lhe foi atribuída de 2 (na escala de 3 a -3). Já no caso da sua associação a 'avisos' (cf. (27)), por força do significado lexical do nome, assume uma polaridade negativa, com intensidade -2.

(26) Os problemas existem e têm de ser resolvidos, haja a coragem política de os reconhecer, mas também de os expor de **forma séria** e direta à UE, procurando zelar pelos interesses dos cidadãos portugueses e não pelos índices de popularidade junto destes. [texto 9]

(27) Desta vez, a OCDE poupou-nos a raspanetes. Mas deixou **avisos sérios**. Manter a santa paz das rendas e privilégios é parar no tempo. E permanecer no atraso que faz de Portugal o país mais injusto da Europa. [texto 2]

O estudo exploratório de natureza linguística descrito consistiu, portanto, na anotação dos valores de sentimento de um conjunto de artigos de opinião da área da economia e finanças, permitindo não só a análise do contributo dos adjetivos e do nome para a expressão da orientação do sentimento ou polaridade (positiva, negativa ou neutra), mas também da sua intensidade (3 a -3), e ainda a elaboração de um léxico de adjetivos, nomes e multipalavras relevantes para a criação de um léxico computacional

específico de economia e finanças, o EconoLex. Além disso, esta análise usou uma metodologia replicável a um conjunto mais vasto de textos (45 no total) ao qual se aplicou uma análise de tipo computacional, com os objetivos, a metodologia e os resultados apresentados na secção 3.3.

4.2. Previsão computacional de valores de sentimento

Nesta secção, pretende-se explorar e conhecer a possibilidade efetiva de um sistema automático poder prever corretamente a polaridade de uma frase, bem como a intensidade do sentimento presente. Seguimos uma abordagem baseada em léxico, dirigindo especial atenção a textos de opinião no domínio da economia e finanças.

4.2.1 Resultados quantitativos

Os resultados deste estudo são apresentados na Tabela 10. Assim, o melhor resultado é obtido com a combinação de léxicos EconoLex⊕SentiLex, pois o valor de EAM é o mais baixo de todos.

TABELA 9 - Valores de EAM resultantes da aplicação de três léxicos de sentimento.

Léxico	EAM
SentiLex	1,53
EconoLex	1,41
EconoLex⊕SentiLex	1,38

O uso exclusivo do EconoLex apresenta também um valor de erro inferior ao do uso exclusivo do léxico SentiLex, mas, na realidade, a junção dos dois atinge o menor erro. Todavia, importa verificar, em termos estatísticos, se esta diferença é ou não significativa.

De modo a verificar a significância estatística dos resultados obtidos, realizou-se o teste não-paramétrico de Wilcoxon com nível de confiança de 95%. Este teste não-paramétrico classifica os valores absolutos das diferenças entre as observações emparelhadas para cada frase e o respetivo léxico

utilizado, calculando assim uma estatística sobre o número de diferenças negativas e positivas.

O *p-value* resultante foi de 0,027 para o EconoLex⊕SentiLex *versus* SentiLex. Como o *valor* foi inferior a 0,05, a diferença entre as observações pode ser considerada estatisticamente significativa. Um teste semelhante foi realizado também para o EconoLex *versus* SentiLex, mas, neste caso, o *p-value* era superior (0.057), isto é, o resultado não é estatisticamente significativo.

4.2.2. Discussão dos resultados

Neste trabalho, comparámos o efeito da introdução de um léxico específico de sentimento (EconoLex) do domínio da economia e finanças na caracterização automática de sentimento, partindo de um léxico de sentimento mais geral (SentiLex). O EconoLex foi especificamente criado a partir de um número modesto de artigos de opinião do domínio da economia e finanças. Seguiu-se, portanto, uma abordagem de análise de sentimento baseada em léxico extraído de um corpus específico de um domínio. Procedeu-se a uma medição do *erro absoluto médio* (EAM), numa avaliação cruzada dos dados e para três configurações de léxicos de sentimento (SentiLex, EconoLex, EconoLex⊕SentiLex).

Os resultados mostraram uma melhoria, embora não muito expressiva, do EconoLex⊕SentiLex em relação ao SentiLex. Na comparação dos resultados EAM com o teste Wilcoxon, constata-se um *p-value* de 0,027, o que dá evidência estatística suficiente para afirmar que o acréscimo de léxico específico (EconoLex) ao léxico geral melhorou o desempenho.

O estudo empírico aqui apresentado poderá ser melhorado com a inclusão de um maior número de textos do domínio. Além disso, poderá também explorar-se a possibilidade de expansão automática do léxico de sentimento deste domínio seguindo/adaptando alguns trabalhos recentes nesta área (Almatarneh & Gamallo, 2018).

5. Considerações finais

Este trabalho teve como objetivos construir um léxico de sentimento específico do domínio da economia e finanças, partindo da avaliação da relevância de nomes, adjetivos e multipalavras na orientação e intensidade dos segmentos de opinião ocorrentes em artigos de opinião, e aferir a eficácia desse léxico na anotação automática do sentimento.

Os resultados da análise do corpus evidenciam a necessidade de melhorar não só o EconoLex, mas também a análise linguística. Em termos de dados, é fundamental alargar o número de textos, idealmente para algumas centenas, bem como proceder a um aumento do léxico específico do domínio, que continua a ser reduzido, relativamente ao SentiLex. Embora não se pretenda que tenha a mesma dimensão, uma vez que se trata de um domínio específico, há, no entanto, um maior leque de termos, expressões multipalavra e construções linguísticas específicas, inerentes ao domínio que nos interessa considerar.

Relativamente ao aumento do léxico de sentimento, no domínio económico, existem diversas possibilidades promissoras para a sua execução de forma automática. Uma abordagem mais clássica consiste no uso de um *thesaurus* (e.g. WordNet-PT; cf. Marrafa 2004), de modo a identificar sinónimos de termos já conhecidos e expandir em vários incrementos sucessivos os termos originais, resultando assim num léxico mais rico.

Uma outra abordagem, mais atual, é usar os chamados *word embeddings*, tais como *word2vec* (Mikolov *et al*. 2013) e *GloVe* (Pennington *et al*. 2014), para fazer o mesmo. A grande diferença neste contexto é que as relações semânticas entre os termos são induzidas automaticamente a partir de corpora, envolvendo o treino de uma rede neuronal multicamada. Uma vantagem evidente desta abordagem é a possibilidade de treinar os modelos com textos específicos de um domínio. No nosso caso, com um bom volume de textos do domínio da economia e finanças, poder-se-ão induzir relações semânticas específicas para estes domínios, menos comuns nos modelos gerais.

Uma outra possibilidade seria introduzir regras semelhantes às que são usadas no tratamento de expressões polares, intensificadores e atenuadores no trabalho de Polanyi & Zaenen (2006) ou de Forte & Brazdil (2016).

No que se refere à análise linguística, estudos como este são prova de

que uma investigação linguística mais detalhada de classes gramaticais como os adjetivos e nomes é necessária, não só para analisar as suas propriedades lexicais na base, mas também as propriedades composicionais na constituição de SN, SADJ. Para além disso, é importante considerar outros elementos/estruturas linguísticas que são determinantes para a caracterização do sentimento em SN/frase, nomeadamente verbos, advérbios e a negação, porque a expressão de sentimento no texto recorre a outros elementos de natureza linguística e discursiva, que não se esgotam nas categorias gramaticais consideradas. Benamara, Taboada & Mathieu (2017) sintetizam esses elementos, dando especial relevo à ocorrência de determinados advérbios, por exemplo, intensificadores e de negação, que podem fazer mudar a polaridade contextual decorrente da consideração de dados lexicais apenas, ou ainda a modalidade, quantificação, relações retóricas, etc.

Tendo em consideração os aspetos que, de acordo com esta investigação, necessitam de melhoramento, é nosso propósito, na sua continuidade, contribuir para o avanço do conhecimento na área da análise de sentimento na interface entre linguística e ciências da computação, com o alargamento do EconoLex e a melhoria dos métodos automáticos de identificação de sentimento em textos de opinião da área da Economia em Português Europeu, em articulação com os dados resultantes da investigação linguística.

REFERÊNCIAS

Abalada, S., Cabarrão, V. & Cardoso, A. 2010. Proposta de Classificação Semântica de Unidades Lexicais Multipalavra Nominais. In: *Textos Seleccionados do XXV Encontro Nacional da APL*. Porto, 81-94.

Adam, J-M. 1997. Unités rédactionnelles et genres discursifs : cadre général pour une approche de la presse écrite. *Pratiques*, 94: 3-18.

Almatarneh, S. & Gamallo, P. 2018. A Comparative Study of Polarity Lexicons to Identify Extreme Opinions. In: *Proceedings of SNAMS 2018, Fifth International Conference on Social Networks Analysis, the Second International Workshop on Advances in*

Natural Language Processing (ANLP 2018) Management and Security. Valencia, Spain, 296-301.

Antunes, P. 2015. *Sentiment Analysis in Financial News*. Dissertação de Mestrado. Porto: FEP.

Baccianella, S., Esuli, A. & Sebastiani, F. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In: *Proceedings of the Seventh conference on International Language Resources and Evaluation*. Valletta, Malta, 2200-2204.

Benamara, F., Taboada, M. & Mathieu, Y. 2017. Evaluative Language Beyond Bags of Words: Linguistic Insights and Computational Applications. *Computational Linguistics*. 43(1): 201-264.

Biber D, Finegan E. 1989. Styles of stance in English: lexical and grammatical marking of evidentiality and affect. *Text*. 9: 93–124.

Calzolari, N. , Fillmore, C. J., Grishman, R., Ide, N., Lenci, A. , MacLeod, C. & Zampolli, A. 2002. Towards Best Practice for Multiword Expressions in Computational Lexicons. In: *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*. Las Palmas, Canary Islands, 1934-1940.

Cambria, E. & Hussain, A. 2015. *Sentic computing: a common-sense-based framework for concept-level sentiment analysis*. Vol. 1. Dordrecht: Springer.

Carvalho, P. & Silva, M.J. 2015. Sentilex-pt: principais características e potencialidades. In: A. Simões, A. Barreiro, D. Santos, R. Sousa-Silva & S.E.O. Tagnin (Eds.). *Linguística, Informática e Tradução: Mundos que se Cruzam*. *Oslo Studies in Language*. 7(1): 425–438.

Carvalho, P., Sarmento, L., Silva, M. J., Oliveira, E. 2009. 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*. Hong Kong, China, 53-56.

Charaudeau, P. 2006. Discours journalistique et positionnements énonciatifs. Frontières et derives. *Semen, 22. 1-9.* Retirado, a 20 de maio de 2017, da Internet: https:// journals.openedition.org/semen/2793.

Cunha, G. X. 2012. A articulação discursiva do gênero artigo de opinião à luz de um modelo modular de análise do discurso. *Filologia Linguística Portuguesa*. 14(1): 73-97.

Das, S. & Chen, M. 2001. Yahoo! for Amazon: Extracting market sentiment from stock message boards. In: *Proceedings of the 8th Asia Pacific Finance Association Annual*

Conference (APFA 2001), Bangkok, Thailand.

Dave, K., Lawrence, S. & Pennock, D.M. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: *Proceedings of International Conference on World Wide Web (WWW-2003)*.

Ekman, P. 1999. Basic emotions. In: T. Dalgleish & M. Power (Eds.). *Handbook of cognition and emotion*. Chichester: John Wiley & Sons, 45-60.

Ferreira, I. 2013. *Para o estudo semântico dos adjetivos adverbiais temporais e aspetuais do Português Europeu*. Tese de Doutoramento. Porto: FLUP.

Fiorin, J.L. 2007. Paixões, afetos, emoções e sentimentos. *CASA: Cadernos de Semiótica Aplicada*. 5 (2), 1-15. Retirado, a 20 de setembro de 2017, da Internet: file:///C:/Users/Fatima%20Silva/Downloads/541-1486-1-PB%20(1).pdf.

Forte, A.C. & Brazdil, P. 2016. Determining the Level of Clients' Dissatisfaction from Their Commentaries. In: J. Silva, R. Ribeiro, P. Quaresma, A. Adami & A. Branco (Eds.). *Computational Processing of the Portuguese Language. PROPOR 2016. Lecture Notes in Computer Science*. Vol. 9727. New York: Springer, 74-85.

Freitas, C. (2013). Sobre a construção de um léxico da afetividade para o processamento computacional do português. *Revista Brasileira de Linguística*, 13 (4), 1031-1059.

Goldberg, A. & Zhu, J. 2006. Seeing stars when there aren't many stars: Graph-based semisupervised learning for sentiment categorization. In *TextGraphs: HLT/NAACL Workshop on Graph-based Algorithms for Natural Language Processing,* 45-52.

Gómez Molina, J. R. 2004. Las unidades léxicas en español. *Carabela*, 56: 27-50.

Hung, C., & Lin, H-K. 2013. Using objective words in SentiWordNet to improve sentiment classification for word of mouth. *IEEE Intelligent Systems*. 28(2): 47-54.

Hunston S, Thompson G. 2000. Evaluation: an introduction. In: S. Hunston & G. Thompson (Eds.). *Evaluation in Text: Authorial Distance and the Construction of Discourse*, Oxford: OUP. 1–27.

Kodratoff, Y. & Michalski, R.S. 2014. *Machine learning: an artificial intelligence approach*. Vol. 3. Massachusetts: Morgan Kaufmann.

Levin, B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.

Li, G. & Liu, F. 2012. Application of a clustering method on sentiment analysis. *Journal of Information Science*. 38(2), 127–139.

Liu, B. 2012. *Sentiment Analysis and Opinion Mining: Synthesis Lectures on Human Language Technologies*. California: Morgan & Claypool Publishers.

Liu, B. 2015. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge:

Cambridge University Press.

Marques-Lucena, M., Sarraipa, J., Fonseca, J., Grilo, A., Jardim-Gonçalves, R. 2015. Framework for customers' sentiment analysis. In: P. Angelov, K.T. Atanassov, L. Doukovska, M. Hadjiski, V. Jotsov, J. Kacprzyk, N. Kasabov, S. Sotirov, E. Szmidt, S. Zadrożny (Eds.). *Intelligent ystems'2014. Advances in Intelligent Systems and Computing*. Vol 322. Cham: Springer, 849-860.

Marrafa, P. 2004. Extending WordNets to Implicit Information. In: M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa & R. Silva (Eds.). *Proceedings of LREC 2004 - International Conference on Language Resources and Evaluation*. Paris: ELRA - European Language Resources Association, 1135-1138 (CD-ROM).

Martin, J.R. & White, P.R.R. (2005). *The Language of Evaluation*. New York: Palgrave.

Mathieu, Y. 2005. Annotation of Emotions and Feelings in Texts. In: J. Tao, T. Tan & R. W. Picard (Eds.). *Affective Computing and Intelligent Interaction. First International Conference, ACII 2005, Beijing, China, October 22-24, 2005, Proceedings*, 350-357.

McCarthy, M. 1990. *Vocabulary*. Oxford: Oxford University Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. & Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In: C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K.Q. Weinberger (Eds.). *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, 3111-3119.

Nasukawa, T. & Yi, J. 2003. Sentiment analysis: Capturing favorability using natural language processing. In: *Proceedings of the 2nd international conference on Knowledge capture*. New York: ACM, 70-77.

Nattinger, J. R., & DeCarrico, J. S. 1992. *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Neviarouskaya, A., Prendinger, H. & Ishizuka, M. (2009). Semantically distinct verb classes involved in sentiment analysis. In: Weghorn, H. & Isaías, P. (Eds.). *Proceedings of the IADIS International Conference on Applied Computing,* 27-34. Retirado, a 28 junho de 2017, da Internet: http://www.iadisportal.org/applied-computing-2009-proceedings.

Pang, B. & Lee, L. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*. 2(1-2): 1-135. Retirado, a 28 junho de 2017, da Internet: http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf.

Pang, B., Lee, L., & Vaithyanathan, S. 2002. Thumbs up?: Sentiment classification using machine learning techniques. *EMNLP '02 - Proceedings of the ACL-02 conference on Empirical methods in natural language processing*. Vol.10, 79-86.

Pennington, J., Socher, R. & Christopher, M. 2014. Glove: Global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing*. Stroudsburg: The Association for Computational Linguistics, 1532-1543.

Polanyi, L. & Zaenen, A. 2006. Contextual valence shifters. In: J.G. Shanahan, Y. Qu, Yan & J. Wiebe (Eds.). *Computing attitude and affect in text: Theory and applications*. Dordrecht: Springer, 1-10.

Ranchhod, E. M. M. 2003. O Lugar das Expressões Fixas na Gramática do Português. In: I. Castro & I. Duarte (Eds.). *Razões e Emoção. Miscelânea de estudos oferecida a Maria Helena Mira Mateus*. Lisboa: Imprensa Nacional - Casa da Moeda, 239-254.

Ravi, K. & Ravi, V. 2015. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*. 89: 14-46.

Rodrigues, R. H. 2005. Os gêneros do discurso na perspectiva dialógica da linguagem: a abordagem de Bakhtin. In: J. L. Meurer, A. Bonini & D. M. Roth (Eds.). *Gêneros: teorias, métodos, debates*. São Paulo: Parábola Editorial, 154–183.

Russell, J.A. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*. 39: 1161-1178.

Sag, I.A., Baldwin, T., Bond, F., Copestake, A. & Flickinger D. 2002. Multiword Expressions: A Pain in the Neck for NLP. In: Gelbukh A. (Ed.) *Computational Linguistics and Intelligent Text Processing. CICLing 2002*. Lecture Notes in Computer Science.Vol. 2276. Berlin: Springer, 189-206.

Silva, F, Leal, A., Silvano, P., Ferreira, I. & Oliveira, F. 2018. Crítica cinematográfica: análise linguístico-textual. In: J. Veloso, P. Silvano, J. Guimarães & R. Sousa e Silva (Eds.). *A linguística em diálogo: volume comemorativo dos 40 anos do Centro de Linguística da Universidade do Porto*. Porto: FLUP / CLUP, 431-458.

Silva, F., Leal, A., Ferreira, I., Oliveira, F. & Silvano, P. 2015. Marcas linguísticas no texto de apreciação crítica. In *Literatura e Gramática: um diálogo infinito*. Lisboa: Associação de Professores de Português.

Silva, M. J. & Team, R. 2011. *Notas sobre a realização e qualidade do twitómetro. Technical report*. Lisboa: FCUL/LASIGE.

Silva, M. J., Carvalho, P. & Sarmento, L. 2012. Building a Sentiment Lexicon for Social Judgement Mining. In: H. Caseli, A. Villavicencio, A. Teixeira & F. Perdigão. (Eds.). *International Conference on Computational Processing of the Portuguese Language (PROPOR)*. Berlin: Springer, 218-228.

Silva, M. J., Carvalho, P., Costa, C. & Sarmento, L. 2010. *Automatic Expansion of a Social*

Judgment Lexicon for Sentiment Analysis. Relatório Técnico: DI–FCUL–TR–2010–08. Lisboa: FCUL.

Taboada, M. & Trnavac, R. 2013. *Nonveridicality and Evaluation Theoretical, Computational and Corpus Approaches*. Brill Academic.

Taboada, M. 2016. Sentiment Analysis: An Overview from Linguistics. *Annual Review of Linguistics 2016*. 2(1): 325-347.

Taboada, M., Anthony, & Voll, K. 2006. Methods for creating semantic orientation dictionaries. In: N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk & D. Tapias (Eds.). *Proceedings of Fifth International Conference on Language Resources and Evaluation (LREC 2006)*. Genoa, Italy, 427-432.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. & Stede, M. 2011. Lexicon-Based Methods for Sentiment Analysis. *Association for Computational Linguistics*. 37(2): 267-307.

Thornbury, S. 2007. *How to teach vocabulary*. Malaysia: Pearson, Longman.

Tong, R. M. 2001. An operational system for detecting and tracking opinions in on-line discussion. In: *Working Notes of the SIGIR Workshop on Operational Text Classification*. New Orleans, 1-6.

Turney, P. D. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, 417-424.

Witten, I. H. & Frank, E. 2016. *Data Mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann.