Realização de Previsões com Conteúdos Textuais em Português

Making Predictions with Textual Contents in Portuguese

Indira Mascarenhas Brito Bruno Martins
Instituto Superior Técnico, INESC-ID
{indira.brito,bruno.g.martins}@tecnico.ulisboa.pt

Resumo

A previsão de quantidades do mundo real com base em informação textual atraiu recentemente um interesse significativo, embora os estudos anteriores se tenham concentrado em aplicações que envolvem apenas textos em inglês. Este artigo apresenta um estudo experimental sobre a realização de previsões com base em textos em português, envolvendo o uso de documentos associados a três domínios distintos. Relatamos experiências utilizando diferentes tipos de modelos de regressão, usando esquemas de ponderação para as caraterísticas descritivas do atual estado da arte, e usando caraterísticas descritivas derivadas de representações para as palavras baseadas no agrupamento automático das mesmas. Através de experiências, demonstramos que modelos de regressão usando a informação textual atingem melhores resultados, quando comparados com abordagens simples tais como realizar as previsões com base no valor médio dos dados de treino. Demonstramos ainda que as representações de documentos mais ricas (e.g., usando o algoritmo de Brown para o agrupamento automático de palavras, e o esquema de ponderação das caraterísticas denominado Delta-TF-IDF) resultam em ligeiras melhorias no desempenho.

Palavras chave

Previsões com Base em Textos, Modelos de Regressão, Agrupamento Automático de Palavras, Engenharia de Caraterísticas em Aplicações de PLN

Abstract

Forecasting real-world quantities, from information on textual descriptions, has recently attracted significant interest as a research problem, although previous studies have focused on applications involving only the English language. This paper presents an experimental study on the subject of making predictions with textual contents in Portuguese, using documents from three distinct domains. We specifically report on experiments using different types of regression models, using state-of-the-art feature weighting schemes, and using features derived from cluster-based word representation. Our experiments show that regression models using the textual information achieve better results than simple baselines such as the average value in the training data, and that richer document representations (i.e., using Brown clusters and the Delta-BM25 feature weighting scheme) results in slight performance improvements.

Keywords

Text-Driven Forecasting, Learning Regression Models, Word Clustering, Feature Engineering for NLP

1 Introdução

A realização de previsões com base em textos atraiu recentemente um interesse significativo nas comunidades internacionais de Extração de Informação, Recuperação de Informação, Aprendizagem Automática, e Processamento de Língua Natural (Smith, 2010; Radinsky, 2012). Exemplos bem conhecidos de estudos anteriores incluem o uso de conteúdos textuais para fazer previsões sobre o comportamento de mercados financeiros (Luo, Zhang e Duan, 2013; Lerman et al., 2008; Tirunillai e Tellis, 2012; Schumaker e Chen, 2009; Bollen, Mao e Zeng, 2011), os resultados de mercados de apostas desportivas (Hong e Skiena, 2010), padrões de vendas de produtos e serviços (Chahuneau et al., 2012; Joshi et al., 2010), eleições governamentais, atividades legislativas e inclinações políticas no geral (Yano, Smith e Wilkerson, 2012; Dahllöf, 2012), ou sondagens de opinião pública em diversos temas (Mitchell et al., 2013; O'Connory et al., 2010; Schwartz et al., 2013).

Este trabalho apresenta um estudo experimental no âmbito da realização de previsões com base em textos em português, usando documentos de três domínios distintos, nomeadamente (i) descrições de hotéis em Portugal recolhidos desde um portal Web[1], associadas aos preços médios dos quartos nas épocas altas e baixas para os turistas, (ii) descrições de restaurantes

[1] http://www.lifecooler.com

e os menus correspondentes, também recolhidos do mesmo portal Web, associados aos preços médios das refeições, e (iii) comentários sobre filmes recolhidos a partir de um site Web especializado[2], em conjunto com os respetivos resultados de bilheteira para a primeira semana de exibição, tal como disponibilizados pelo Instituto do Cinema e do Audiovisual[3]. O nosso estudo incidiu sobre o uso de métodos de aprendizagem automática do atual estado-da-arte (e.g., regressão com florestas aleatórias, ou regressão linear com regularização dada pela método da rede elástica), implementados numa biblioteca Python para aprendizagem automática, chamada *scikit-learn*[4]. Além da questão dos conteúdos em português, o nosso estudo também apresenta algumas novidades técnicas em relação à maior parte do trabalho anterior na área, nomeadamente através de (i) experiências com esquemas de ponderação de caraterísticas do atual estado-da-arte, como o Delta-TF-IDF ou o Delta-BM25, e (ii) experiências com caraterísticas derivadas de representações com base no algoritmo de Brown para o agrupamento automático de palavras.

O restante conteúdo do artigo está organizado da seguinte forma: a Seção 2 apresenta trabalhos anteriores relevantes. A Seção 3 detalha as principais contribuições do artigo, apresentando as técnicas de regressão que foram consideradas, bem como as abordagens para a representação dos conteúdos textuais como vetores de caraterísticas descritivas. A Seção 4 apresenta a avaliação dos métodos propostos, descrevendo os conjuntos de dados dos três domínios diferentes, e discutindo os resultados obtidos. Finalmente, a Seção 5 apresenta as principais conclusões e aponta possíveis direções para trabalho futuro.

2 Trabalhos Relacionados

Inspirando-se em trabalhos recentes sobre análise de sentimentos (Pang e Lee, 2008), em que técnicas de aprendizagem automática são usadas para interpretar textos com base na atitude subjetiva dos autores, Noah Smith e os seus colegas têm abordado várias tarefas de prospeção de texto relacionadas, onde os documentos textuais são interpretados para prever os valores de variáveis de interesse do mundo real. Este é um dos grupos de investigadores com mais atividade nesta área específica. Um artigo relativamente recente, que resume o trabalho que estes investigadores têm vindo a desenvolver, foi publicado on-line por Smith (2010). Exemplos específicos para as tarefas de previsão com base em textos, que estes autores abordaram, incluem:

1. A interpretação de relatórios financeiros anuais, publicados por empresas aos seus acionistas, a fim de tentar prever o risco incorrido ao investir numa empresa, no ano seguinte (Kogan et al., 2009);
2. A interpretação de comentários sobre filmes, feitos por críticos de cinema, com o objetivo de tentar prever o resultado de bilheteira dos filmes (Joshi et al., 2010);
3. Interpretar mensagens colocadas em blogues políticos, para tentar prever a resposta recolhida dos leitores (Yano e Smith, 2010);
4. A interpretação de mensagens diárias em *microblogs*, a fim de prever opinião pública e a confiança dos consumidores (Yano, Cohen e Smith, 2009; Yano e Smith, 2010);
5. Interpretar textos correspondentes a descrições de restaurantes, menus de restaurantes, e comentários de clientes, para tentar prever o preço médio de refeições e as avaliações dos clientes (Chahuneau et al., 2012).

Em todos os casos acima, um aspeto do significado do texto é observável a partir de dados objetivos do mundo real, embora talvez não imediatamente no momento em que o texto é publicado (ou seja, respetivamente, observa-se a volatilidade dos retornos, a receita bruta, os comentários dos utilizadores, resultados de questionários de opinião tradicionais, e os preços médios das refeições, nos cinco problemas que foram anteriormente enumerados). Smith (2010) propôs uma abordagem genérica para a previsão baseada em textos, suportada em modelos de regressão que utilizam caraterísticas descritivas derivadas do texto, as quais são geralmente ruidosas e esparsas. Este autor argumentou que a previsão com base em textos, como um problema de investigação, pode ser abordada por meio de metodologias baseadas em aprendizagem que são neutras a diferentes teorias linguísticas (Smith, 2010).

Por exemplo no que diz respeito às críticas e receitas de filmes, e sumarizando o trabalho anterior de Joshi et al. (2010), Smith mencionou que antes da estreia de um filme, os críticos assistem e publicam comentários sobre o mesmo. Os autores procuraram realizar previsões sobre os resultados de bilheteira com base nos comentários, à medida que estes são produzidos pelos críticos. Consideraram-se 1,351 filmes lançados entre Janeiro de 2005 e Junho de 2009. Para cada filme, foram obtidos dois tipos de dados:

[2] `http://www.portal-cinema.com`
[3] `http://www.ica-ip.pt`
[4] `http://scikit-learn.org`

1. Metadados descritivos recolhidas a partir do site *Metacritic*[5], incluindo o nome, a produtora, o(s) gênero(s), realizador(es), diretor(es), os atores principais, e o país de origem, entre outros dados. Metadados de um site chamado *The Numbers*[6] foram também recolhidos, contendo informação sobre o orçamento de produção, as receitas brutas do final de semana da estreia do filme, e o número de salas de cinema em que o filme foi exibido nesse final de semana.
2. Comentários extraídos dos seis sites de análise de filmes mais referenciados no site *Metacritic*, e apenas comentários publicados antes da data de estreia do filme.

Os autores descrevem a aplicação de um modelo de regressão linear com regularização dada pelo método da rede elástica (Zou e Hastie, 2005; Fridman, Hastie e Tibshirani, 2008). O modelo foi treinado em 988 exemplos lançados entre 2005 e 2007, e foi avaliado na previsão da receita de bilheteira para cada filme lançado entre Setembro de 2008 e Junho de 2009 (i.e., um total de 180 filmes). Foi calculado o erro absoluto médio (MAE) sobre o conjunto de teste, analisando a diferença entre a receita estimada para cada filme durante a sua semana de lançamento, e os ganhos brutos reais, por ecrã. Modelos que usam apenas o texto (MAE de 6,729$), ou o texto em adição a metadados (MAE de 6,725$), foram melhores do que os modelos que usam apenas os metadados (MAE de 7,313$). O texto reduz o erro por 8% em relação aos metadados, e por 5% quando comparado com uma forte base de previsão dos resultados de bilheteira, dada pelo valor médio dos filmes nos de dados de treino.

Mais recentemente, Chahuneau et al. (2012) exploraram as interações no uso da linguagem que ocorrem entre os preços dos menus de restaurantes, descrições de itens do menu, e sentimentos expressos em comentários de utilizadores, a partir de dados extraídos do site *Allmenus*[7]. Deste site, os autores recolheram menus de restaurantes em sete cidades norte-americanas, nomeadamente *Boston*, *Chicago*, *São Francisco*, *Los Angels*, *Nova Iorque*, *Washington DC* e *Filadélfia*. Cada menu contém uma lista de nomes de itens, com descrições textuais opcionais e preços. Metadados adicionais (e.g., gama de preço, localização, e ambiente) e comentários dos clientes (i.e., descrições textuais associadas a uma classificação numa escala de 0 a 5 estrelas), para a maioria dos restaurantes, foram recolhidos a partir de um conhecido site Web chamado *Yelp*[8].

Os autores consideraram diversas tarefas de previsão, tais como a previsão de preços individuais de itens, a previsão da gama de preços para cada restaurante, e a previsão em conjunto do preço médio e da opinião dos clientes. Para as duas primeiras tarefas, os autores utilizaram modelos de regressão linear, e para a terceira tarefa, usaram modelos de regressão logística, em todos os casos com regularização l_1. Para a avaliação, os autores usaram métricas como o erro absoluto médio (MAE) ou o erro relativo médio (MRE).

Para prever o preço individual de cada item num menu, Chahuneau et al. (2012) utilizaram o logaritmo do preço como o valor a modelar e a prever, pois a distribuição dos preços é mais simétrica numa escala logarítmica. Os autores avaliaram várias estratégias simples que fazem previsões independentes para cada nome diferente dos itens nos menus. Duas destas estratégias usam a média e a mediana do preço, no conjunto de treino e dado o nome do item. Uma terceira estratégia utiliza um modelo de regressão linear com regularização l_1, que foi treinado com múltiplas caraterísticas binárias. Os autores realizaram uma simples normalização dos nomes de itens em todas estas estratégias, devido à sua grande variação no conjunto de dados (i.e., mais de 400 mil nomes distintos). A normalização consiste na remoção das palavras mais frequentes nos nomes dos itens, ordenando de seguida as palavras em cada nome lexicograficamente. Esta normalização reduziu o número de nomes por 40%.

Os autores exploraram diferentes modelos ricos em caraterísticas descritivas, com base em regressão regularizada, considerando (i) características binárias para cada propriedade de metadados do restaurante, (ii) n-gramas em nomes de itens do menu, em que n-gramas correspondem a sequências de n palavras (i.e., com $n \in \{1, 2, 3\}$) extraídas de um determinado nome, (iii) n-gramas nas descrições de itens do menu, e (iv) n-gramas de menções a itens do menu nos comentários correspondentes. Ao usar o conjunto completo de caraterísticas descritivas, os autores relatam uma redução final de 0,5 na métrica MAE, e de cerca de 10% no MRE. Os autores reportam assim bons resultados para esta técnica, quando comparados com os resultados das estratégias elementares.

Na tarefa de prever a gama de preços, os valores alvo eram números inteiros de 1 a 4 que indicam o custo de uma refeição típica do restaurante. Na avaliação desta tarefa, os autores

[5] `http://www.metacritic.com`
[6] `http://www.the-numbers.com`
[7] `http://www.allmenus.com`
[8] `http://www.yelp.com`

arredondaram os valores previstos para inteiros, e usaram o erro absoluto médio (MAE) e a exatidão, como medidas da qualidade dos resultados. Os autores notaram uma pequena melhoria ao comparar o modelo de regressão linear com um modelo de regressão ordinal (i.e., um modelo que atribui, para cada caso, um valor de classificação entre 1 e 4, e que leva em consideração a ordenação dos valores alvo (McCullagh, 1980)), medindo 77,32% de exatidão na regressão ordinal, contra 77,15%, para modelos com os metadados. Os autores também usaram caraterísticas descritivas do texto completo dos comentários, além das caraterísticas utilizadas para a tarefa de previsão de preços individuais de itens do menu. Ao combinar os metadados e as caraterísticas descritivas dos comentários, a exatidão medida excedeu o valor de 80%.

Para a tarefa de analisar as opiniões expressas nos comentários, os autores treinaram um modelo de regressão logística, prevendo a polaridade da opinião expressa em cada comentário. A polaridade de um comentário foi determinada pela correspondente pontuação de classificação em estrelas, ou seja, se está acima ou abaixo da média. A exatidão obtida foi de 87%.

Finalmente, Chahuneau et al. (2012) consideraram a tarefa de prever, em conjunto, o preço médio e a opinião agregada para um restaurante. Para fazer isso, os autores tentaram modelar, ao mesmo tempo, a polaridade dos comentários $\overline{r}$ e o preço dos itens $\overline{p}$. Os autores calcularam, para cada restaurante no conjunto de dados, a média do preço dos itens e a média da pontuação em estrelas. Um plano ($\overline{r}$, $\overline{p}$) foi dividido em quatro seções, com os pontos médios dos dois atributos no conjunto de dados como as coordenadas de origem, ou seja, 8,69$ para $\overline{p}$ e 3,55 estrelas para $\overline{r}$. Esta divisão permitiu aos autores treinar um modelo de regressão logística de 4 classes, usando as caraterísticas descritivas extraídas das avaliações para cada restaurante. A exatidão obtida foi, neste caso, de 65%.

3 Realização de Previsões com Base em Conteúdos Textuais

Neste estudo, à semelhança do trabalho anterior de Noah Smith e seus colegas, abordámos o problema de fazer previsões, com conteúdos textuais, como uma tarefa de regressão. Reportamos resultados para experiências em três domínios distintos, com textos escritos em português, e considerando algumas inovações, tais como o uso de agrupamentos de palavras ou de diferentes esquemas de ponderação das caraterísticas.

Cada documento é modelado como um vetor de caraterísticas descritivas num determinado espaço vetorial, em que a dimensionalidade corresponde ao número de caraterísticas descritivas diferentes. Esta representação está associada a um modelo bem conhecido para o processamento e representação de documentos na área da recuperação de informação, normalmente referido como o modelo do espaço vetorial. Formalmente, tem-se que cada documento é representado como um vetor de caraterísticas descritivas $\overrightarrow{d_j} =< w_{1,j}, w_{2,j}, ..., w_{k,j} >$, em que k é o número de caraterísticas, e em que cada $w_{i,j}$ corresponde a um peso que reflete a importância da caraterística i para a descrição dos conteúdos do documento j. As caraterísticas descritivas são, essencialmente, as palavras que ocorrem na coleção de documentos. No entanto, em algumas das nossas experiências, também usámos outras caraterísticas, tais como propriedades de metadados referentes à localização geográfica (i.e., os distritos administrativos) associados aos exemplos, os tipos de restaurantes, ou agrupamentos de palavras associados aos termos que ocorrem no documento correspondente.

3.1 Agrupamento Automático de Palavras

O agrupamento automático de palavras semelhantes permite abordar o problema da esparsidade dos dados, ao proporcionar uma representação de dimensionalidade inferior para as palavras de uma coleção de documentos. Neste trabalho, foi utilizado o algoritmo de agrupamento de palavras proposto por Brown et al. (1992), que induz representações generalizadas de palavras individuais. Este algoritmo é essencialmente um processo de agrupamento hierárquico que forma grupos de palavras com caraterísticas comuns, a fim de maximizar a informação mútua de bi-gramas. A entrada para o algoritmo é um corpus de texto, que pode ser visto como uma sequência de N palavras $w_1, \cdots, w_N$. O resultado é uma árvore binária, na qual as folhas são as palavras. O processo de agrupamento está relacionado com um modelo de linguagem baseado em bi-gramas e classes:

$$\mathrm{P}(w_1^N|C) = \prod_{i=1}^{N} \mathrm{P}\left(\mathrm{C}(w_i)|\mathrm{C}(w_{i-1})\right) \times \mathrm{P}\left(w_i|\mathrm{C}(w_i)\right)$$

Na fórmula, $\mathrm{P}(c|c')$ corresponde a probabilidade de transição da classe c dada a sua classe antecessora c', e $\mathrm{P}(w|c)$ é a probabilidade de emissão da palavra w numa dada classe c. As probabilidades do modelo podem ser calculadas por

contagem de frequências relativas de unigramas e bi-gramas. Para determinar as classes ótimas C para um número de classes M, podemos adotar uma abordagem de máxima verosimilhança do tipo $\mathrm{C} = \arg\max_{\mathrm{C}} \mathrm{P}(W_1^N|\mathrm{C})$. Brown et al. (1992) demonstraram que o melhor agrupamento é o que resulta da maximização da informação mútua entre as classes adjacentes, dada por:

$$\sum_{c,c'} \mathrm{P}(c,c') \times \log\left(\frac{\mathrm{P}(c,c')}{\mathrm{P}(c) \times \mathrm{P}(c')}\right)$$

A estimativa do modelo de linguagem é, portanto, baseada num procedimento de agrupamento automático aglomerativo, que é usado para construir uma estrutura hierárquica sobre as distribuições de contextos de cada palavra. O algoritmo começa com um conjunto de nós folha, um para cada uma das classes de palavras (i.e., inicialmente, uma classe para cada uma das palavras). Em seguida, de forma iterativa, são selecionados pares de nós a fundir, otimizando de forma gananciosa um critério de qualidade baseado na informação mútua entre agrupamentos adjacentes (Brown et al., 1992). Cada palavra é, portanto, inicialmente atribuída ao seu próprio grupo/classe, e o algoritmo funde pares de classes, de modo a induzir a redução mínima na informação mútua, parando quando o número de classes é reduzido para o número predefinido $|C|$.

Neste trabalho, para induzir representações para as palavras, utilizámos uma implementação *open source*[9] do algoritmo de Brown, que seque a descrição dada por Turian, Ratinov e Bengio (2010). Este *software* foi usado em conjunto com uma grande coleção de textos escritos em português, à qual tínhamos acesso e que representa diversos tipos de temas e géneros linguísticos. Este textos correspondem a um conjunto de frases que combina o corpus CINTIL do português moderno (Barreto et al., 2006), com artigos noticiosos publicados no jornal *Público*[10], durante um período de 10 anos. Induzimos um total de mil grupos de palavras, onde cada grupo tem um identificador único a usar nas representações.

3.2 Ponderação das Caraterísticas

Neste trabalho, experimentámos diferentes formas de calcular o peso das caraterísticas descritivas a serem usadas nos nossos modelos, para os termos textuais e os agrupamentos de palavras semelhantes. Estas formas incluem o uso de valores binários, a frequência dos termos, TF-IDF, e também esquemas de ponderação mais sofisticados, tais como os esquemas Delta-TF-IDF e Delta-BM25 discutidos por Martineau e Finin (2009) e por Paltoglou e Thelwall (2010).

No caso de pesos binários, cada parâmetro $w_{i,j}$ toma o valor zero ou um, dependendo se o elemento i está presente ou não no documento j.

Outra abordagem comum é usar a frequência de ocorrência de cada elemento i no documento j, tendo-se que é comum considerar uma penalização logarítmica para estes valores.

TF-IDF é, talvez, o esquema de ponderação de caraterísticas mais popular, combinando a frequência individual de cada elemento i num documento j (i.e., a componente *Term Frequency*, ou TF), com a frequência inversa do elemento i na coleção de documentos (i.e., a componente *Inverse Document Frequency*, ou IDF). Quando $n_i > 0$, o peso TF–IDF de um elemento i para um documento j é dado pela seguinte fórmula:

$$\mathrm{TF\text{–}IDF}_{i,j} = \log_2(1 + \mathrm{TF}_{i,j}) \times \log_2\left(\frac{N}{n_i}\right)$$

Na fórmula, N é o número total de documentos na colecção, e n_i é o número dos documentos contendo o elemento i. TF–IDF é zero se $n_i \leq 0$.

Os esquemas de ponderação Delta–TF–IDF e Delta–BM25 medem a importância relativa de um termo em duas classes distintas. No contexto dos nossos problemas de regressão, não temos classificações binárias associadas a cada uma das instâncias, mas sim valores reais. No entanto, considerámos duas classes a fim de determinar os pesos das caraterísticas descritivas de acordo com estes esquemas, ao dividir as instâncias entre aquelas que tem um valor superior ou igual à mediana dos valores nos dados de treino, e aquelas que são menores ou iguais à mediana.

O esquema Delta–TF–IDF estende a fórmula do TF–IDF, localizando a estimação das pontuações de IDF para os documentos associados a cada uma das duas classes, subtraindo depois os dois valores. O peso de um elemento i num documento j pode ser obtido como se mostra na seguinte equação, quando $\mathrm{TF}_{i,j} > 0$:

$$\Delta\mathrm{TF\text{–}IDF}_{i,j} = \log_2(1 + \mathrm{TF}_{i,j}) \times \log_2\left(1 + \frac{N_{pos} \times n_{i,neg}}{n_{i,pos} \times N_{neg}}\right)$$

Cada parâmetro N_c corresponde ao número de documentos de treino na colecção c, e $n_{i,c}$ é o número de documentos da coleção c no qual o termo i ocorre. No contexto deste trabalho, c

[9] `http://github.com/percyliang/brown-cluster`
[10] `http://www.publico.pt`

pode ser *positivo* para os exemplos com o valor superior à mediana, e *negativo* para os casos inferiores à mediana. De acordo com um grande conjunto de experiências relacionadas com a classificação binária de opiniões em textos, a abordagem Delta–TF–IDF é significativamente melhor do que esquemas baseados em TF ou numa ponderação binária (Martineau e Finin, 2009).

Paltoglou e Thelwall (2010) concluíram que ao introduzir, adicionalmente, variantes localizadas e suavizadas das funções IDF, em conjunto com esquemas de ponderação TF escalados, a exatidão pode ser aumentada ainda mais. No esquema de ponderação Delta–BM25, o peso de um elemento i para um documento j é dado pela seguinte fórmula, onde s é uma constante de suavização que é normalmente definida como 0,5:

$$\begin{aligned}\Delta\text{BM25}_{i,j} &= \log_2(1+\text{TF}_{i,j}) \\ &\quad\times \log_2\left(1+\frac{(N_{pos}-n_{i,pos}+s)\times n_{i,neg}+s}{(N_{neg}-n_{i,neg}+s)\times n_{i,pos}+s}\right)\end{aligned}$$

3.3 Modelos de Regressão

Poderiam ter sido usados vários tipos de modelos de regressão, a fim de abordar as tarefas de previsão com base em textos que foram por nós consideradas. Neste trabalho, comparámos modelos de regressão linear, usando diferentes tipos de regularização, com modelos baseados na combinação de várias árvores de regressão.

3.3.1 Métodos de Regressão Linear

Considerando um conjunto de dados $\{y_i, x_{i1}, ..., x_{ik}\}_{i=1}^{n}$ com n instâncias, e assumindo que a relação entre a variável dependente y_i e um vetor de k caraterísticas descritivas x_i é linear, tem-se que uma regressão linear assume a seguinte forma:

$$\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{pmatrix} \times \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \cdots \\ b_k \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \cdots \\ e_n \end{pmatrix}$$

Na fórmula, $x_{i,j}$ corresponde à *i–ésima* caraterística descritiva do *j–ésimo* exemplo, os parâmetros b_i correspondem aos coeficientes de regressão, e e_j é um erro que capta a diferença entre a forma efetiva das respostas observadas y_i, e os resultados da previsão do modelo de regressão. A fórmula pode ser re-escrita usando uma notação matricial, ficando $y = Xb + e$.

Vários procedimentos têm sido desenvolvidos para a estimativa de parâmetros em modelos de regressão linear. O método dos mínimos quadrados (i.e., *linear least squares regression*, ou LSR) é a forma mais simples e mais utilizada para estimar os parâmetros desconhecidos num modelo de regressão linear. O método LSR minimiza a soma dos quadrados dos resíduos $\text{S}(b)$ entre os dados e o modelo, ou seja, minimiza a soma $\sum_{i=1}^{n} e_i^2$. O quadrado dos resíduos pode ser re-escrito em notação matricial como $e'e$, onde o apóstrofo significa que a matriz foi transposta. Substituindo e por $y - Xb$, temos que:

$$\begin{aligned}\text{S}(b) &= \sum_{i=1}^{n} e_i^2 \\ &= (y-Xb)'(y-Xb) \\ &= y'y - y'Xb - b'X'y + b'X'Xb\end{aligned}$$

A condição para a $\text{S}(b)$ estar no mínimo é ter as derivadas $\frac{\partial \text{S}(b)}{\partial b} = 0$. O primeiro termo da equação acima não depende de b, enquanto que o segundo e o terceiro termos são iguais nas suas derivadas, e o último termo é uma forma quadrática dos elementos b. Assim, temos que:

$$\frac{\partial \text{S}(b)}{\partial b} = -2X'y + 2X'Xb$$

Igualando a equação diferencial a zero, temos:

$$\begin{aligned}-2X'y + 2X'Xb &= 0 \\ X'Xb &= X'y \\ b &= (X'X)^{-1}X'y \\ b &= \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) \\ b &= \arg\min_b ||y-Xb||^2\end{aligned}$$

Na regressão com o método dos mínimos quadrados, é bem conhecido que quando o número de instâncias de treino n for menor do que o número de caraterísticas k, ou se existirem muitas caraterísticas correlacionadas, os coeficientes de regressão podem apresentar uma alta variação, e tendem a ser instáveis. Neste caso, são necessários métodos de regularização para melhor ajustar o modelo aos dados, e para manter a variação dos coeficientes de regressão sob controlo.

A abordagem da regressão em crista (i.e., *ridge regression*) penaliza o tamanho dos coeficientes, por adição de uma penalização l_2 correspondente a um termo $||b||_2^2$ no modelo. Assim, os

coeficientes de regressão são estimados através do seguinte problema de otimização:

$$b = \arg\min_b ||y - Xb||^2 + \lambda||b||_2^2$$

Na fórmula, $\lambda \geq 0$ é um parâmetro de ajuste, o qual controla a força do termo de regularização. Quando $\lambda = 0$ obtemos a estimativa de regressão linear regular, e quando $\lambda = \infty$ obtemos $b = 0$.

Tibshirani (1996) propôs um outro método de regularização, chamado *Least Absolute Shrinkage and Selection Operator* (Lasso), que encolhe alguns coeficientes e fixa outros a zero. O método Lasso utiliza uma penalização l_1 dada por $||b||_1$ como forma de regularização, e tenta estimar os parâmetros b através do seguinte problema:

$$b = \arg\min_b ||y - Xb||^2 + \lambda||b||_1$$

Uma das principais diferenças entre o Lasso e a regressão em crista é que na regressão em crista, quando a penalização é aumentada, todos os parâmetros são reduzidos mas permanecem ainda diferentes de zero, enquanto que com a regularização Lasso, aumentar a penalização conduz alguns dos parâmetros a zero. Assim, os modelos Lasso tendem a ser esparsos, na medida em que utilizam menos caraterísticas descritivas. No entanto, uma limitação do Lasso é que se $k > n$, então este método seleciona no máximo n variáveis, ou seja, o número de variáveis selecionadas é limitado pelo número de exemplos de treino. Outra limitação do método Lasso ocorre quando existe um grupo de variáveis altamente correlacionadas. Neste caso, o Lasso tende a selecionar uma variável apenas a partir deste grupo, ignorando as outras. Para resolver estes problemas, Zou e Hastie (2005) propuseram a abordagem de regularização conhecida como o método da rede elástica (i.e., *elastic net*), a qual combina as regularizações l_1 e l_2 com pesos λ_1 e λ_2, respetivamente. As estimativas deste método são definidas pela seguinte fórmula:

$$b = \arg\min_b ||y - Xb||^2 + \lambda_1||b||_1 + \lambda_2||b||_2^2$$

O método da rede elástica tende a produzir melhores estimativas quando as variáveis são consideravelmente correlacionadas. O método elimina a limitação do número de variáveis selecionadas, incentiva efeitos de agrupamento, e estabiliza a abordagem de regularização l_1.

Vários métodos têm sido propostos para estimar modelos de regressão linear com regularização *ridge*, Lasso, e *elastic net*, incluindo a pesquisa cíclica por coordenadas no sentido descendente do gradiente (i.e., *cyclical coordinate descent*) (Kim e Kim, 2006), ou outros métodos de otimização convexa com base em cálculos iterativos (Boyd e Vandenberghe, 2004), tais como o SpaRSA (Wright, Nowak e Figueiredo, 2009). Neste estudo, foi utilizada a implementação disponível a partir do pacote Python de aprendizagem automática denominado *scikit-learn*.

3.3.2 Métodos de Aprendizagem por Combinação

Os métodos de aprendizagem por combinação (i.e., *ensemble learning*) tentam combinar simultaneamente vários modelos, que são geralmente modelos simples baseados em árvores de decisão, para obter um melhor desempenho em problemas de previsão (Sewell, 2011). Neste trabalho, foram utilizados dois tipos de métodos de aprendizagem por combinação, nomeadamente o método da floresta aleatória (i.e., *random forest*) de árvores de regressão, e um modelo de regressão baseado em múltiplas árvores de regressão obtidas por reforço em relação ao gradiente de uma dada função (i.e., *gradient boosted regression trees*), mais uma vez tal como implementados no pacote Python denominado *scikit-learn*.

A floresta aleatória de árvores de regressão combina a ideia de *bagging*, desenvolvida por Breiman (1996), com uma seleção aleatória de caraterísticas descritivas (Breiman, 2001). Em suma, temos que este método constrói uma coleção de árvores de-correlacionadas, e produz como resultado o seu valor médio. O principal objetivo deste método é reduzir a variância no modelo final de regressão, através da redução da correlação entre as variáveis. Isto é conseguido pela seleção aleatória de variáveis. Seja N o número de casos de treino, e seja M o número de instâncias utilizadas para treino do modelo. O algoritmo procede da seguinte forma:

1. Escolher um conjunto de treino para cada árvore, amostrando n vezes com substituição de todos os N casos de treino disponíveis. Usar as instâncias restantes para estimar o erro da árvore, ao fazer as previsões.

2. Para cada nó da árvore, selecionar m variáveis aleatoriamente para apoiar a decisão naquele nó, e calcular a melhor divisão para a árvore com base nessas m variáveis e no conjunto de treino da etapa anterior.

3. Cada árvore cresce até atingir a maior extensão possível, e nenhuma poda é realizada. O algoritmo CART tal como proposto

por Breiman et al. (1984) é utilizado para a geração das várias árvores.

Os passos acima são iterados para gerar um conjunto de árvores. Ao fazer uma previsão, a média das previsões de todas as árvores é relatada. Cada árvore vota com um peso correspondente ao seu desempenho no subconjunto de dados que foi deixado de parte durante o treino.

Quanto ao método *gradient boosted regression trees* (GBRT), este é por sua vez baseado na ideia de *boosting*, suportando funções de otimização específicas para problemas de regressão, como a soma dos erros quadráticos (Friedman, 2001). Um modelo GBRT consiste de um conjunto de modelos fracos, normalmente árvores de decisão (i.e., árvores CART, semelhantes às que são utilizadas no caso de regressão com florestas aleatórias). O modelo toma a seguinte forma, onde $\mathrm{h}_m(X)$ são as funções de base do modelo:

$$y = \mathrm{F}(X) = \sum_{m=1}^{M} \mathrm{h}_m(X)$$

De acordo com o princípio da minimização do risco empírico, o método tenta minimizar o valor médio de uma função de perda em relação ao conjunto de treino. Isto faz-se começando com um modelo, que consiste de uma função constante F_0, e incrementalmente expandindo-o de forma gananciosa, de acordo com a seguinte equação:

$$\mathrm{F}_m = \mathrm{F}_{m-1}(X) - \gamma_m \mathrm{h}_m(X)$$

Em cada estágio, uma árvore de decisão $\mathrm{h}_m(X)$ é escolhida para minimizar a função de perda considerada (i.e., a soma dos quadrados dos erros), dado o modelo atual F_{m-1} e as suas previsões $\mathrm{F}_{m-1}(X)$ (i.e., as árvores h_m aprendidas em cada etapa são geradas usando pseudo-resíduos como o conjunto de treino). A inicialização para o modelo F_0 é escolhida frequentemente com base na média dos valores alvo, e o multiplicador γ_m é encontrado em cada fase, resolvendo o seguinte problema de otimização, onde L é a função de perda considerada:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} \mathrm{L}(y_i, \mathrm{F}_{m-1}(x_i) + \gamma \mathrm{h}_m(x_i))$$

O problema de otimização é normalmente abordado através de um procedimento de descida ao longo do gradiente (i.e., *gradient descent*) da função (Boyd e Vandenberghe, 2004).

4 Validação Experimental

Este trabalho explorou a realização de previsões com base em textos de três domínios distintos com caraterísticas diferentes, ou seja, considerando informação sobre hotéis, restaurantes e os seus preços em Portugal, ou sobre comentários de filmes, em conjunto com os resultados de bilheteira correspondentes. Para os hotéis e os restaurantes, recolhemos descrições textuais do site Lifecooler[11]. Para cada restaurante, a informação disponibilizada por este site inclui o nome, a descrição textual, o menu, as especialidades, o tipo de restaurante, o preço médio de uma refeição, e a localização, que inclui o nome da cidade e o respetivo distrito. Para os hotéis, a informação disponível inclui o nome, a descrição textual, a localização, e o preço dos quartos nas épocas altas e baixas. Para o caso dos filmes, usamos comentários do site Portal do Cinema[12], e metadados provenientes do Instituto do Cinema e do Audiovisual[13], para os filmes que foram lançados entre 2008 e 2013 em Portugal. A informação disponível inclui o nome do filme, o distribuidor, o produtor, o número de salas em que o filme foi exibido durante a semana de estreia, a receita bruta da primeira semana, o número de espectadores, e uma classificação por estrelas numa escala de 0 a 5. Considerámos apenas os filmes encontrados nos dois sites, e assim acabámos por utilizar um conjunto de 502 filmes. As principais caraterísticas descritivas, dos conjuntos de dados associados aos três domínios, estão apresentadas na Tabela 1. Para os hotéis e os restaurantes, os valores alvo são mostrados em Euros, enquanto que para o caso dos filmes são apresentados valores em milhares de Euros.

Os conjuntos de dados diferem em vários aspetos, tais como no número de documentos disponíveis e na distribuição dos valores a serem previstos. A distribuição dos valores alvo para os três domínios é apresentada na Figura 1.

Todo o texto disponível foi utilizado nas nossas experiências (e.g., para os hotéis, utilizámos o nome do hotel e a descrição textual, enquanto que para os restaurantes utilizámos o nome do restaurante, a descrição textual, as especialidades, e os menus. Para os filmes, usámos o nome do filme e o comentário). Para além das caraterísticas textuais, em algumas experiências utilizámos a localização, para hotéis e restaurantes, os tipos dos restaurantes, ou o número de salas em que o filme foi exibido. A localização (i.e., os

[11] `http://www.lifecooler.com`
[12] `http://www.portal-cinema.com`
[13] `http://www.ica-ip.pt`

	Hotéis Alta	Hotéis Baixa	Restaurantes	Filmes
Número de descrições textuais	2656	2656	4677	502
Tamanho do vocabulário	9932	9932	19421	28720
Número médio de termos por texto	35	35	47	346
Valor alvo mínimo	10,00	10,00	4,50	0,93
Valor alvo máximo	3000,00	1200,00	100,00	1437,71
Média dos valores alvo	95,92	71,48	18,10	162,75
Mediana dos valores alvo	75,00	60,00	15,00	73,56
Desvio padrão nos valores alvo	93,42	51,67	8,41	229,21

Tabela 1: Caraterização estatística dos três conjuntos de dados que foram considerados.

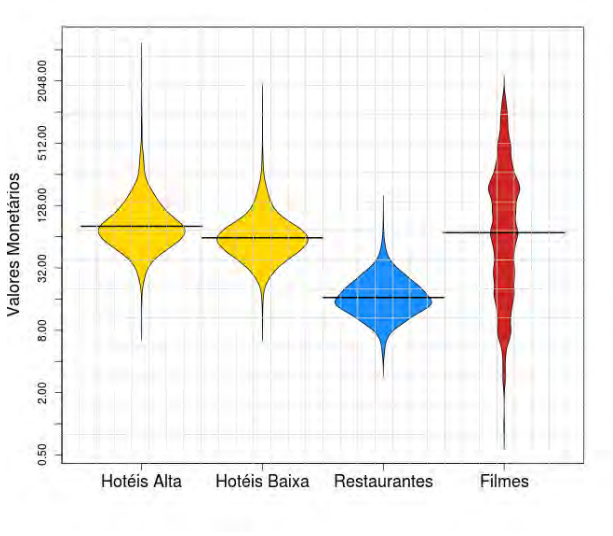

Figura 1: Distribuição dos valores a serem previstos, nos vários conjuntos de dados.

distritos administrativos) pode naturalmente influenciar quão caro é um hotel ou um restaurante. Por exemplo, como podemos ver na Figura 2, os distritos com os hotéis e os restaurantes mais caros são Lisboa e Faro, e estes mesmos distritos são os que apresentam a maior variação de preços.

4.1 Metodologia Experimental

Todas as experiências foram realizadas com uma metodologia de validação cruzada usando 10 desdobramentos (i.e., *folds*). A qualidade dos resultados foi aferida usando métricas como o erro absoluto médio (i.e., o *mean absolute error*) e o desvio padrão empírico generalizado (i.e., a métrica *root mean squared error*).

O erro absoluto médio (MAE) é uma medida que compara as previsões com os valores reais, e que corresponde a uma média dos valores de erro absoluto, como mostra a Equação 1. O desvio padrão empírico generalizado é outra medida da exatidão de modelos de previsão, calculada com base na raiz quadrada da média dos quadrados dos erros, como mostrado na Equação 2.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

Considerando um conjunto de dados $\{y_i, x_{i1}, ..., x_{ik}\}_{i=1}^{n}$, onde os valores x_{ik} correspondem às caraterísticas, onde y_i corresponde aos verdadeiros valores alvo, e tendo $\hat{y}_i$ como os resultados previstos, pode-se facilmente ver que as métricas anteriores estimam erros nas mesmas unidades de medida que os valores alvo, ou seja, em Euros ou em milhares de Euros, no caso das experiências relatadas neste trabalho.

Além das métricas MAE e RMSE, reportamos também alguns resultados em termos de uma variante normalizada do erro médio, correspondendo ao erro relativo médio (i.e., *mean relative error*) tal como apresentado na equação abaixo:

$$\text{MRE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{3}$$

O erro relativo médio (MRE) permite estabelecer comparações entre tarefas de previsão diferentes, dado que esta medida é independente das unidades e da escala em que as variáveis a ser previstas se encontram expressas.

4.2 Resultados Obtidos

Num primeiro teste, tentámos prever os preços dos quartos de hotéis, o preço médio dos pratos em restaurantes, ou as receitas de bilheteira de filmes, usando apenas conteúdos textuais. Nesta tarefa, comparámos modelos de regressão com abordagens simples, tais como realizar as previsões com base no valor médio e na mediana nos dados de treino. Foram consideradas representações baseadas no esquema de ponderação de termos mais popular, ou seja, TF-IDF. Como podemos ver na Tabela 2, os modelos de regressão com caraterísticas derivadas dos textos atingiram

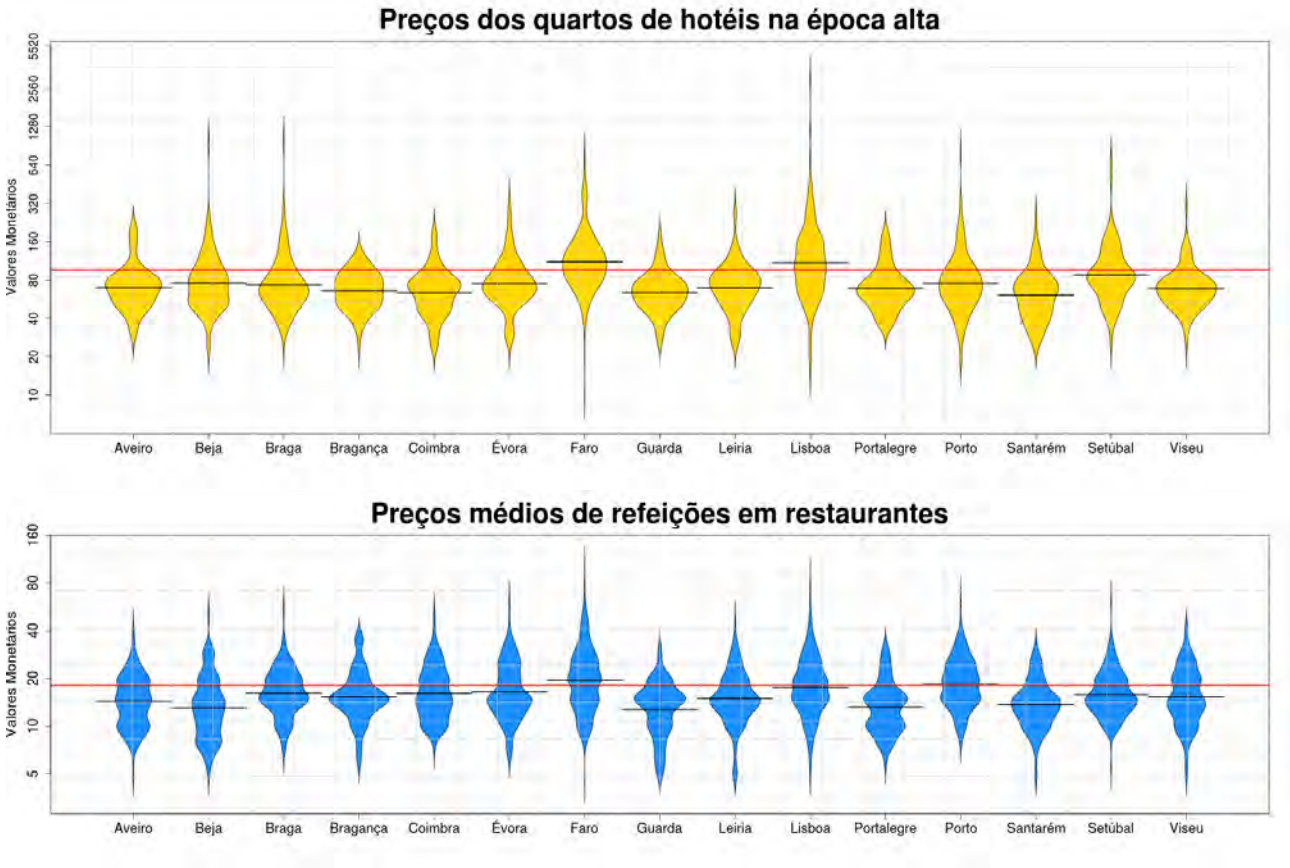

Figura 2: Distribuição para os valores a serem previstos, por distrito em Portugal Continental.

melhores resultados que as abordagens mais simples. Verificamos também que, de todos os modelos considerados, os melhores resultados foram obtidos com o método de regularização da rede elástica. O melhor modelo baseado em aprendizagem por combinação corresponde à abordagem da floresta aleatória de árvores de regressão.

Num conjunto separado de experiências, procurámos analisar a importância das diferentes caraterísticas descritivas correspondentes aos termos, vendo as suas diferenças relativas em termos da contribuição para prever os valores alvo. Isto foi feito para o caso de modelos com base na técnica da floresta aleatória de árvores de regressão, ou com base em regressão linear com regularização dada pelo método da rede elástica, usando pesos para as caraterísticas descritivas calculados com a abordagem TF-IDF.

No caso de modelos de regressão linear com regularização dada pelo método da rede elástica, inspecionamos os pesos das caraterísticas descritivas (i.e., os coeficientes de regressão) dos modelos aprendidos, e calculámos a média dos pesos de cada caraterística sobre as múltiplas *folds* dos nossos testes de validação cruzada. No caso de modelos baseados na técnica da floresta aleatória de árvores de regressão, a posição relativa (i.e., a profundidade) de uma caraterística que seja usada como nó de decisão numa árvore pode servir para avaliar a importância relativa dessa caraterística, no que diz respeito à previsão. As caraterísticas que são utilizadas na parte superior das árvores contribuem para a decisão final de uma maior fração dos exemplos e, deste modo, a fração esperada dos exemplos, para as quais a caraterística contribui, pode ser utilizada como uma estimativa da importância relativa das caraterísticas. Pela média dessas taxas de atividade esperada, ao longo das várias árvores, e pela média também sobre as múltiplas *folds* das experiências de validação cruzada, pode-se estimar uma importância para cada caraterística.

A Figura 3 ilustra as 20 caraterísticas descritivas mais importantes em termos dos coeficientes de regressão linear (i.e., as 10 caraterísticas descritivas com os maiores valores positivos ou negativos), ou em termos da posição relativa dos nós de decisão, para o caso de modelos de previsão de preços de quartos de hotel nas épocas altas e baixas. Como esperado, termos como *sheraton, luxo* ou *resort* parecem indicar preços mais altos, enquanto que termos como *pensão* ou *hostel* estão associados a preços mais baixos.

	Hotéis Alta		Hotéis Baixa		Restaurantes		Filmes	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Média	44,06	93,65	28,58	51,86	6,09	8,41	161,15	247,87
Mediana	39,96	95,69	26,59	52,92	5,80	8,97	140,48	264,29
Regressão em Crista	45,87	72,94	30,38	42,41	6,40	6,83	127,70	201,52
Lasso	35,78	72,96	24,27	43,60	4,59	6,57	183,01	268,33
Rede Elástica	**34,63**	**70,86**	**23,25**	**41,97**	**4,27**	**6,20**	**127,55**	**192,70**
Floresta Aleatória	**34,25**	**74,13**	**23,17**	**44,25**	**4,40**	**6,56**	**135,89**	**211,77**
Gradient Boosting	37,91	79,94	25,18	47,09	4,65	7,02	166,74	269,95

Tabela 2: Resultados da primeira experiência, com uma representação baseada em TF-IDF.

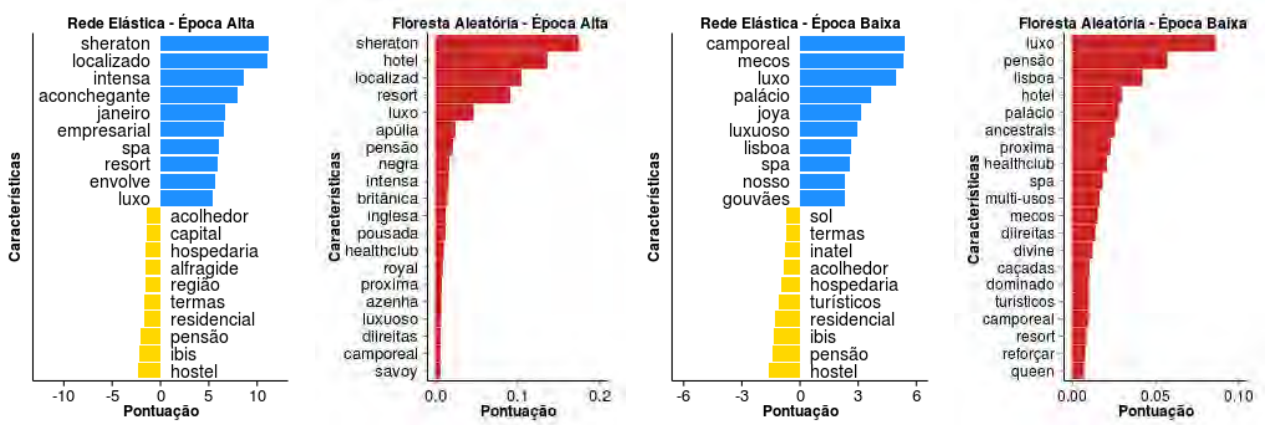

Figura 3: As 20 caraterísticas mais importantes para a previsão dos preços de quartos de hotéis.

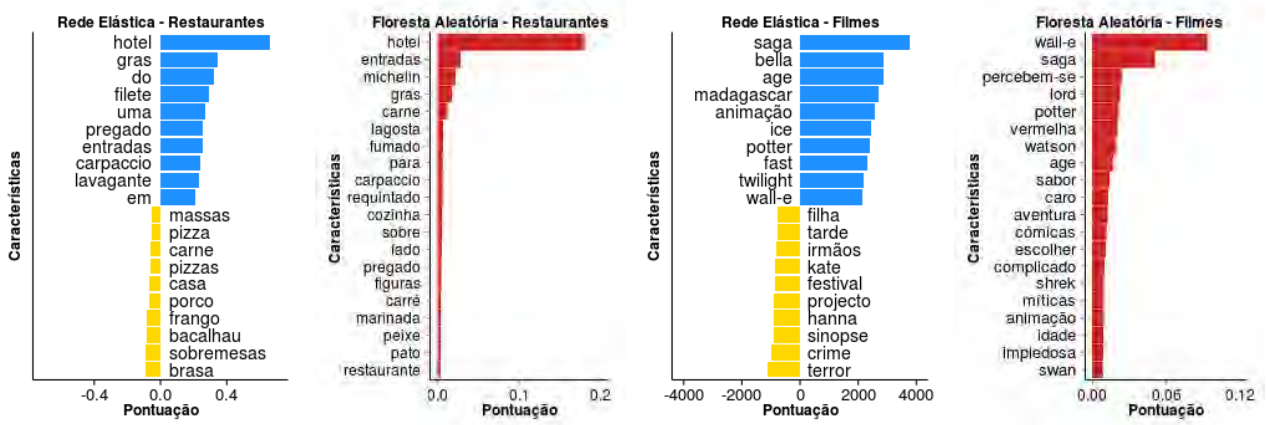

Figura 4: As 20 caraterísticas mais importantes para a previsão dos preços médios dos pratos em restaurantes, ou para a previsão dos resultados de bilheteira de filmes.

A Figura 4 mostra as 20 caraterísticas descritivas mais importantes, mas neste caso para os modelos de previsão de preços de refeições em restaurantes (i.e., os gráficos do lado esquerdo), e para os modelos de previsão de resultados de bilheteira de filmes. No caso de restaurantes, termos como *hotel* ou *michelin* parecem ser muito discriminativos, enquanto que no caso dos filmes, termos como *saga* ou *terror* parecem fornecer as melhores pistas. Na Figura 4 é possível notar que algumas palavras correspondentes a *stop-words* (e.g., as palavras *do* ou *uma*), ou correspondentes a nomes de filmes específicos (e.g., *wall-e*) estão associadas a importâncias altas nos modelos de regressão. Nos nossos testes, não usámos nenhuma estratégia de remoção das *stop-words*, e temos que estes valores podem indicar efeitos de sobre-ajustamento (i.e., *over-fitting*) dos modelos aos dados de treino. Importa realçar que os testes foram realizados apenas com conjuntos de dados relativamente pequenos.

As Tabelas 3 e 4 mostram uma comparação dos resultados obtidos com os melhores modelos, ou seja, a regressão linear com regularização com o método da rede elástica, e regressão com florestas aleatórias, respetivamente, utilizando cada

	Hoteis Alta		**Hotéis Baixa**		**Restaurantes**		**Filmes**	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Binário	40,91	77,49	27,14	46,64	5,94	8,22	133,01	199,34
Frequência do Termo	51,18	86,10	30,34	48,64	5,42	6,31	209,50	279,51
TF-IDF	34,63	70,86	23,25	41,97	4,27	6,20	127,55	192,70
Delta-TF-IDF	**34,55**	**70,63**	24,33	41,77	4,36	6,62	131,59	194,37
Delta-BM25	34,70	72,82	**23,21**	**40,24**	**4,22**	**6,14**	**127,41**	**191,08**

Tabela 3: Resultados com modelos de regressão com regularização dada pelo método da rede elástica, usando diferentes esquemas de ponderação das caraterísticas.

	Hotéis Alta		**Hotéis Baixa**		**Restaurantes**		**Filmes**	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Binário	36,56	79,06	24,28	46,19	4,83	7,08	135,68	214,60
Frequência do Termo	36,62	77,09	25,53	46,69	5,37	6,57	137,92	207,82
TF-IDF	34,25	74,13	**23,17**	44,25	4,40	6,56	135,89	211,77
Delta-TF-IDF	**34,12**	**73,43**	24,91	44,04	**4,32**	6,85	**130,59**	**209,49**
Delta-BM25	34,47	73,55	23,19	**43,45**	4,63	**6,52**	134,84	210,24

Tabela 4: Resultados com modelos baseados em florestas aleatórias de árvores de regressão, usando diferentes esquemas de ponderação das caraterísticas.

uma das representações para os conteúdos textuais. A representação mais rica é, talvez, dada pelo esquema Delta-BM25, embora o esquema Delta-TF-IDF tenha obtido resultados muito semelhantes ou até melhores, no caso dos modelos aprendidos com base na abordagem da floresta aleatória de árvores de regressão.

Além dos conteúdos textuais, considerámos caraterísticas derivadas de outros elementos de metadados, tais como a localização geográfica de hotéis e restaurantes, o tipo de restaurante, ou o número de salas em que o filme foi exibido no fim de semana de estreia. Também experimentámos medir os resultados após a adição de caraterísticas derivadas de agrupamentos automáticos de palavras semelhantes (i.e., *word clusters* gerados com o algoritmo de Brown), para a representação dos conteúdos textuais.

As propriedades de metadados, tais como a localização ou o tipo de restaurante, são representadas como valores binários (por exemplo, uma caraterística para cada distrito administrativo), assumindo o valor de um ou zero, dependendo da localização ou do tipo correspondente à descrição textual. Para o caso dos filmes, adicionámos ainda o número de salas como uma das dimensões do vetor de caraterísticas.

A Tabela 5 lista os resultados correspondentes à previsão de preços de quartos de hotéis com diferentes conjuntos de caraterísticas descritivas. A combinação de conteúdos textuais, metadados e agrupamentos de palavras obteve um melhor desempenho, no caso do uso da abordagem de regularização da rede elástica. Ao utilizar a técnica da floresta aleatória de árvores de regressão, a combinação de texto e localização geográfica atingiu melhores resultados.

A Tabela 6 apresenta os resultados para a previsão do preço médio das refeições em restaurantes, usando diferentes conjuntos de caraterísticas. Com o método da rede elástica, a experiência que considera apenas as caraterísticas descritivas baseadas nas palavras produziu melhores resultados. Com a abordagem da floresta aleatória de árvores de regressão, a combinação de caraterísticas derivadas do texto e localização atingiu melhores resultados.

Finalmente, a Tabela 7 mostra os resultados para a previsão de receitas de bilheteira de filmes. O número de salas tem uma forte influência nos resultados. Com ambos os tipos de modelos de regressão, a execução que obteve melhores resultados é claramente aquela que envolve a combinação de caraterísticas derivadas do texto com o número de salas. O número de salas e as receitas de bilheteira são, de facto, altamente correlacionadas, como mostra a Figura 5.

Para restaurantes e hotéis, também comparámos os resultados obtidos através dos nossos modelos de regressão contra abordagens mais simples, como prever com base no valor médio dos dados de treino, ou com base no valor médio e mediano por localização. No caso dos filmes, tentámos também usar apenas o número de salas. Estes resultados estão apresentados na Tabela 8. Conseguimos melhores resultados ao fazer as previsões com base no valor médio por localização, comparando com o uso da média ao longo de todo o conjunto de dados de treino. Na Tabela 8, apresentamos também os resultados em termos da

		Texto		+WClusters		+Localização		Todas	
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Época Baixa	**Rede Elástica**	23,21	40,24	25,52	45,06	23,57	43,05	**23,17**	**40,19**
	Floresta Aleatória	23,19	43,45	25,33	46,06	**23,18**	**43,06**	23,76	44,23
Época Alta	**Rede Elástica**	34,70	72,82	39,39	75,47	34,75	70,46	**34,00**	**70,13**
	Floresta Aleatória	34,47	73,55	38,87	77,99	34,38	76,46	**34,02**	**73,30**

Tabela 5: Resultados da previsão de preços de quartos em hotéis, com diferentes conjuntos de características descritivas.

	Texto		+WClusters		+Tipo		+Localização		Todas	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Rede Elástica	**4,22**	**6,14**	5,52	7,82	4,88	7,03	4,87	7,02	4,71	6,79
Floresta Aleatória	4,63	6,52	5,11	7,55	4,36	6,59	**4,33**	**6,52**	4,34	6,44

Tabela 6: Resultados da previsão de preços de refeições em restaurantes, com diferentes conjuntos de caraterísticas descritivas.

	Texto		+WClusters		+Salas		Todas	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Rede Elástica	127,91	191,08	126,48	191,30	79,06	125,45	**72,95**	**122,29**
Floresta Aleatória	127,41	191,08	138,46	220,45	**66,02**	**137,78**	72,49	135,14

Tabela 7: Resultados da previsão para os resultados de bilheteira associados a filmes, com diferentes conjuntos de caraterísticas descritivas.

métrica MRE, por forma a suportar comparação entre diferentes tarefas. A previsão de receitas de bilheteira de filmes apresenta-se como um problema mais difícil, com resultados ligeiramente piores em termos da medida MRE. Este problema em concreto é também aquele onde temos um menor volume de dados disponíveis.

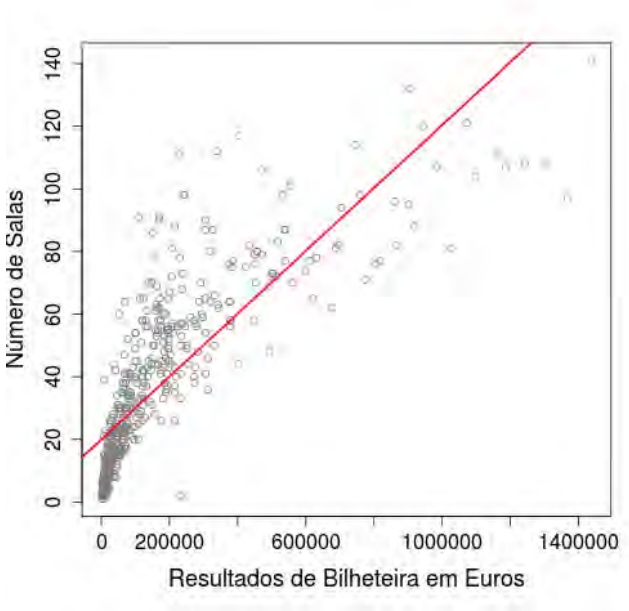

Figura 5: Resultados de bilheteira versus o número de salas no qual o filme foi apresentado.

Em suma, e como relatado na Tabela 8, podemos concluir que os modelos de regressão que usam caraterísticas derivadas do conteúdo textual de fato resultam em ganhos de exatidão para as tarefas de previsão consideradas. A adição de caraterísticas descritivas dos metadados, tais como a localização, resulta apenas em ligeiras melhorias sob modelos de regressão que usam caraterísticas baseadas no texto.

5 Conclusões e Trabalho Futuro

Este artigo apresentou um estudo experimental sobre a realização de previsões com base em conteúdos textuais escritos em português, e usando documentos associados a três domínios diferentes. As tarefas específicas abordadas no nosso trabalho envolveram (i) a previsão de preços de quartos para hotéis em Portugal, nas épocas altas e baixas para os turistas, usando descrições textuais recolhidas a partir de um portal Web conhecido, (ii) a previsão de preços médios de refeições em restaurantes localizados em Portugal, com descrições textuais para os restaurantes e os seus menus, recolhidos também a partir do mesmo portal Web, e (iii) a previsão de resultados de bilheteira de filmes, na primeira semana de exibição, conforme relatado pelo Instituto do Cinema e do Audiovisual, para filmes

	Hotéis Alta			Hotéis Baixa			Restaurantes			Filmes		
	MAE	RMSE	MRE	MAE	RMSE	MRE	MAE	RMSE	MRE	MAE	RMSE	MRE
Média	44,06	93,65	0,54	28,58	51,86	0,46	6,09	8,41	0,38	161,15	247,87	5,38
Média por Localização	40,57	90,80	0,48	28,11	50,91	0,46	5,81	8,11	0,36	–	–	–
Mediana por Localização	37,45	92,11	0,38	26,48	51,80	0,37	5,78	8,45	0,33	–	–	–
Número de Salas	–	–	–	–	–	–	–	–	–	83,45	131,89	2.01
Melhor Modelo	**34,00**	**70,13**	**0,37**	**23,17**	**40,19**	**0,35**	**4,22**	**6,14**	**0,31**	**66,02**	**122,29**	**0,63**

Tabela 8: Resultados gerais para as diferentes tarefas de previsão.

exibidos em Portugal e usando comentários textuais de outro portal Web bem conhecido. Relatámos especificamente experiências utilizando diferentes tipos de modelos de regressão, usando esquemas de ponderação para as caraterísticas do actual estado da arte, e usando caraterísticas derivadas de representações para as palavras baseadas no agrupamento automático das mesmas. Através das nossas experiências, conseguimos demonstrar claramente que os modelos de previsão usando a informação textual alcançam melhores resultados do que abordagens mais simples, tais como realizar previsões com base no valor médio dos dados de treino. Demonstrámos ainda que o uso de representações de documentos mais ricas (e.g., usando o algoritmo de Brown para o agrupamento automático de palavras, e o esquema de ponderação das caraterísticas Delta-TF-IDF) resultara em ligeiras melhorias no desempenho.

Apesar dos resultados interessantes, há muitas ideias para trabalho futuro. Dado que só temos acesso a conjuntos de dados de treino relativamente pequenos, acreditamos que um caminho interessante se relaciona com a avaliação de técnicas de aprendizagem semi-supervisionada, capazes de aproveitar grandes quantidades de dados não anotados. Também nos parece razoável supor que as pistas para estimar corretamente um determinado valor alvo, com base num documento textual, podem estar contidas num subconjunto pequeno das frases do documento. Yogatama e Smith (2014) introduziram um algoritmo de aprendizagem que explora esta intuição, através de uma abordagem de regularização cuidadosamente projetada, mostrando que o método resultante pode superar significativamente outras abordagens (e.g., regularizadores padrão como os métodos *ridge*, Lasso, e a rede elástica) em diversos problemas de categorização de texto. Finalmente, dado o sucesso parcial das representações de documentos feitas com base no algoritmo de agrupamento automático de palavras desenvolvido por Brown, gostaríamos de experimentar com outros tipos de representações baseadas em similaridade distribucional, tais como as representações de palavras enquanto vetores densos de baixa dimensionalidade, propostas no estudo de Mikolov et al. (2013).

Agradecimentos

Este trabalho foi parcialmente financiado pela Fundação para a Ciência e a Tecnologia (FCT), através do projeto com referência UTA-Est/MAI/0006/2009 (REACTION), bem como através do financiamento plurianual do laboratório associado INESC-ID, com a referência PEst-OE/EEI/LA0021/2013.

Agradecemos particularmente aos nossos colegas do projeto REACTION acima mencionado, pela sua ajuda e pelas observações pertinentes.

Referências

Barreto, Florbela, António Branco, Eduardo Ferreira, Amália Mendes, Maria Fern, A Bacelar Do Nascimento, Filipe Nunes, e João Ricardo Silva. 2006. Open resources and tools for the shallow processing of Portuguese: The TagShare project. Em *Proceedings of the International Conference on Language Resources and Evaluation*.

Bollen, Johan, Huina Mao, e Xiao-Jun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 1(2).

Boyd, Stephen e Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.

Breiman, Leo. 1996. Bagging Predictors. *Machine Learning*, 24(2).

Breiman, Leo. 2001. Random Forests. *Machine Learning*, 45(1).

Breiman, Leo, J. H. Friedman, R. A. Olshen, e C. J. Stone. 1984. *Classification and Regression Trees*. Wadsworth and Brooks.

Brown, Peter F., Peter V. de Souza, Robert L. Mercer, Vincent J. Della Pietra, e Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4).

Chahuneau, Victor, Kevin Gimpel, Bryan R. Routledge, Lily Scherlis, e Noah A. Smith. 2012. Word salad: Relating food prices and

descriptions. Em *Proceedings of the Conference on Empirical Methods on Natural Language Processing*.

Dahllöf, Mats. 2012. Automatic prediction of gender, political affiliation, and age in swedish politicians from the wording of their speeches—a comparative study of classifiability. *Literary and Linguistic Computing*, 27(2).

Fridman, Jerome, Trevor Hastie, e Rob Tibshirani. 2008. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1).

Friedman, Jerome H. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5).

Hong, Yancheng e Steven Skiena. 2010. The wisdom of bookies? sentiment analysis vs. the NFL point spread. Em *Proceedings of the AAAI International Conference on Weblogs and Social Media*.

Joshi, Mahesh, Dipanjan Das, Kevin Gimpel, e Noah A. Smith. 2010. Movie reviews and revenues: An experiment in text regression. Em *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Kim, Yongdai e Jinseog Kim. 2006. Gradient Lasso for feature selection. Em *Proceedings of the International Conference on Machine Learning*.

Kogan, Shimon, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, e Noah A. Smith. 2009. Predicting risk from financial reports with regression. Em *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Lerman, Kevin, Ari Gilder, Mark Dredze, e Fernando Pereira. 2008. Reading the markets: forecasting public opinion of political candidates by news analysis. Em *Proceedings of the International Conference on Computational Linguistics*.

Luo, Xueming, Jie Zhang, e Wenjing Duan. 2013. Social media and firm equity value. *Information Systems Research*, 24(1).

Martineau, Justin e Tim Finin. 2009. Delta TF-IDF: An improved feature space for sentiment analysis. Em *Proceedings of the AAAI International Conference on Weblogs and Social Media*.

McCullagh, Peter. 1980. Regression models for ordinal data. *Journal of Royal Statistical society*, 42(2).

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, e Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. Em *Proceedings of the Conference on Neural Information Processing Systems*.

Mitchell, Lewis, Morgan R. Frank, Kameron Decker Harris, Peter Sheridan Dodds, e Christopher M. Danforth. 2013. The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLOS ONE*, 8(5).

O'Connory, Brendan, Ramnath Balasubramanyany, Bryan R. Routledge, e Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. Em *Proceedings of the AAAI International Conference on Weblogs and Social Media*.

Paltoglou, Georgios e Mike Thelwall. 2010. A study of information retrieval weighting schemes for sentiment analysis. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Pang, Bo e Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2).

Radinsky, Kira. 2012. Learning to predict the future using Web knowledge and dynamics. *ACM SIGIR Forum*, 46(2).

Schumaker, Robert P. e Hsinchun Chen. 2009. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems*, 27(12).

Schwartz, H. Andrew, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Megha Agrawal Gregory J. Park, Shrinidhi K. Lakshmikanth, Sneha Jha, Martin E. P. Seligman, Lyle Ungar, e Richard E. Lucas. 2013. Characterizing geographic variation in well-being using tweets. Em *Proceeding of the AAAI International Conference on Weblogs and Social Media*.

Sewell, Martin. 2011. Ensemble methods. Relatório Técnico RN/11/02, University College London Departament of Computer Science.

Smith, Noah A. 2010. Text-Driven Forecasting.

Tibshirani, Robert. 1996. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B*, 58(1).

Tirunillai, Seshadri e Gerard J. Tellis. 2012. Does chatter really matter? Dynamics of

User-Generated Content and Stock Performance. *Information Systems Research*, 31(2).

Turian, Joseph, Lev Ratinov, e Yoshua Bengio. 2010. Word representation: a simple and general method for semi-supervised learning. Em *Proceedings of the Annual Meeting of the Association for Computional Linguistics*.

Wright, Stephen J., Robert D. Nowak, e Mário A. T. Figueiredo. 2009. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7).

Yano, Tae, William W. Cohen, e Noah A. Smith. 2009. Predicting response to political blog posts with topic models. Em *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Yano, Tae e Noah A. Smith. 2010. What's worthy of comment? Content and Comment Volume in Political Blogs. Em *Proceedings of the AAAI International Conference on Weblogs and Social Media*.

Yano, Tae, Noah A. Smith, e John D. Wilkerson. 2012. Textual predictors of bill survival in congressional committees. Em *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Yogatama, Dani e Noah A. Smith. 2014. Making the most of bag of words: Sentence regularization with alternating direction method of multipliers. Em *Proceedings of the International Conference on Machine Learning*.

Zou, Hui e Trevor Hastie. 2005. Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society Series B*, 67(5).