10-2019

LexiNB - A two-step approach for sentiment classification on tweets related to Portuguese tax authorities

Alcides de Almeida Seiça

Antonio Trigo

Fernando Paulo Belfo

LexiNB - Uma abordagem bietápica de classificação de sentimentos em *tweets* relacionados com as autoridades fiscais portuguesas

LexiNB - A two-step approach for sentiment classification on tweets related to Portuguese tax authorities

Alcides de Almeida Seiça, Instituto Politécnico de Coimbra, ISCAC, Quinta Agrícola, Bencanta, 3040-316 Coimbra, Portugal, alcidesaseica@gmail.com

António Trigo, Instituto Politécnico de Coimbra, ISCAC, Quinta Agrícola, Bencanta, 3040-316 Coimbra, Portugal / Universidade do Minho, Centro ALGORITMI, 4804-533 Guimarães, Portugal, antonio.trigo@gmail.com

Fernando Paulo Belfo, Instituto Politécnico de Coimbra, ISCAC, Quinta Agrícola, Bencanta, 3040-316 Coimbra, Portugal, fpbelfo@gmail.com

Resumo

A crescente importância das redes sociais na nossa sociedade leva governos e instituições públicas a privilegiarem estas redes, não só, na comunicação com os seus cidadãos, mas também na perceção da opinião e grau de satisfação que os cidadãos têm sobre os serviços prestados. Recorrendo a técnicas de *text mining*, mais especificamente, de análise de sentimentos (*sentiment analysis* ou *opinion mining*), é possível extrair informação útil das redes sociais que permita a identificação e acompanhamento das opiniões dos cidadãos. Neste sentido, apresentamos neste artigo, o trabalho desenvolvido a partir de *tweets* da rede social Twitter, relativos às autoridades fiscais portuguesas, com o objetivo de explorar algoritmos de classificação de texto que permitam identificar as opiniões, positivas e negativas, dos cidadãos. Como principal contributo deste estudo, destacamos a proposta de uma nova abordagem, designada por "LexiNB" que consiste na aplicação do algoritmo *Naive Bayes* a um conjunto de dados gerado a partir da utilização da biblioteca "Lexicon-PT" sobre o conjunto de dados original. Esta nova abordagem apresentou um aumento significativo da estatística *Kappa* e da eficiência *(Balanced Accuracy)*.

Palavras-chave: Autoridade Tributária; *Text Mining*; Análise de Sentimentos; *Naive Bayes*

Abstract

The growing importance of social networks in our society leads governments and public institutions to privilege these networks, not only in communicating with their citizens, but also in the perception of citizens' opinion and degree of satisfaction with the services provided. By using text mining techniques, more specifically, of sentiment analysis or opinion mining, it is possible to extract useful information from social networks that allows the identification and monitoring of citizens' opinions. In this sense, we present in this article, the work developed, from tweets of the social network Twitter related to the Portuguese Tax Authority, with the purpose of exploring algorithms of text classification, that allow to identify the opinions, positive and negative, of the citizens. As a main contribution of this study, we highlight the discovery of a new approach, which consists in applying the Naive Bayes algorithm to a dataset generated from the use of the Lexicon-PT library on the original dataset, which was named LexiNB. This new approach showed a significant increase in the Kappa statistic and Balanced Accuracy.

Keywords: *Tax Authorities; Text Mining; Sentiment Analysis; Naive Bayes*

ISSN 2183-489X

1. INTRODUÇÃO

O rápido crescimento e massificação da utilização da Internet e a enorme popularidade das redes sociais alteraram as formas de interação entre pessoas e organizações. A rede social Twitter surgiu em 2006 e apresentava uma funcionalidade inovadora: cada *post*, geralmente designado por *tweet*, não deve exceder o limite de 140 caracteres. A cada minuto, são publicados cerca de 350 mil *tweets* (Twitter Stats 2018) e os mesmos dispõem de informações valiosas que possibilitam às organizações descobrir a opinião desses utilizadores em relação aos seus produtos e serviços (Filho, 2014)

Os governos utilizam as redes sociais devido à sua enorme popularidade, como forma de facilitação e aumento da interatividade a atratividade da sua comunicação com os cidadãos, tentando colocar um "rosto humano" no governo e potenciando também o alcance de novas audiências. Na nota informativa do Fórum sobre Administração Fiscal da OCDE de 2011, subordinado ao tema; "Social Media Technologies (SMTs)", reconhece-se que o desenvolvimento e a utilização de SMTs se encontra ainda numa fase muito precoce e é útil refletir sobre a sua evolução (Forum on Tax Administration, 2011).

Este estudo tem como objetivo a exploração e identificação de algoritmos no domínio do *text mining*, mais especificamente na tarefa de classificação de textos, que permitam identificar as opiniões, positivas e negativas, dos cidadãos relativamente às autoridades fiscais portuguesas expressas nos *tweets* da rede social Twitter.

Estruturámos este trabalho em seis secções. Após a introdução, nas secções 2 e 3 apresentamos uma revisão da literatura e dos trabalhos relacionados com pertinência para o estudo do caso. A secção 4 apresenta a metodologia de recolha, pré-processamento dos tweets, seleção e classificação manual da amostra. Na secção 5 apresentamos os resultados do processo de extração de características e na secção 6 evidenciamos os resultados das técnicas e modelos escolhidos. Terminamos este estudo com a apresentação, na secção 7, das principais conclusões obtidas e reservámos a seção 8 para agradecimento do apoio concedido pela FCT - Fundação para a Ciência e Tecnologia.

2. REFERENCIAL TEÓRICO

A rápida expansão da internet provocou um aumento exponencial do volume de informação disponível sob a forma de opiniões discutidas em fóruns, comunidades, redes sociais, etc. Indurkhya & Damerau (2010) afirmam que as opiniões são tão importantes que, em qualquer área em que seja necessário tomar decisões, os decisores querem ouvir a opinião de outros.

A realização deste trabalho entronca na ideia de 'sentimento' relacionado com o conceito de 'opinião' enquanto base dos valores pessoais de um indivíduo, os seus sentimentos e emoções.

O processo de classificação de sentimentos com base no seu "valor sentimental" consiste em agrupar as opiniões em duas classes: opiniões positivas e opiniões negativas. Para esse efeito, é criado um

modelo que consiga identificar a classe a que a opinião pertence. Os trabalhos realizados por Pang, Lee, & Vaithyanathan (2002) concretizam a aplicação de técnicas de aprendizagem de máquina neste domínio.

O *Naive Bayes* é um algoritmo que se tornou popular na área de aprendizagem de máquina ("*Machine Learning*") para categorizar textos baseado na frequência das palavras usadas. Trata-se de um classificador suportado no teorema de *Bayes* frequentemente utilizado na classificação de textos por ser rápido e fácil de implementar (Rennie, Shih, Teevan, & Karger, 2003).

$$P(A|B) = P(B|A) \times P(A) / P(B)$$

O teorema de *Bayes* vem representado na fórmula anterior. Sendo B um evento passado e A um evento que depende de B, a probabilidade à posteriori da ocorrência do evento A dado o evento B, representada por P(A|B), é calculada através da contagem dos casos em que A e B ocorrem juntos dividindo pelo número de casos em que apenas ocorre o evento B.

Uma outra técnica de classificação é a baseada em recursos léxicos, também chamada de baseada em dicionários ou linguística e tem como foco o emprego de dicionários de palavras e/ou expressões que possuem classes pré-definidas e valores. A técnica mais utilizada com recursos léxicos é a da ocorrência de sentimento próximo a uma característica da entidade. Esta técnica possui bons resultados quando empregada em frases ou textos pequenos, pois a proximidade com a característica é menor. Quando a análise é feita em textos maiores, as características podem estar numa frase, e a opinião noutra, o que pode reduzir a precisão da técnica (Hu & Liu, 2004).

Os léxicos são bases de dados criadas manualmente (ou de forma automatizada), em que as palavras têm já as respetivas polaridades (positiva = 1, negativa = -1 ou neutra = 0) associadas. São normalmente designados por léxicos de opinião e apresentam-se como um instrumento fundamental em *text mining* (Liu & Zhang, 2012). A vantagem principal dos léxicos é que não é necessário rotular os dados para treino. A sua maior desvantagem é que os léxicos são limitados a um idioma e ao tamanho da base de dados.

3. TRABALHOS RELACIONADOS

Nos últimos anos temos assistido à divulgação de inúmeras experiências que envolvem a utilização de *tweets* para análise de sentimentos recorrendo ao método *Naive Bayes* para uma ampla gama de contextos.

No campo da administração pública surgiram algumas pesquisas sobre a utilização de redes sociais análise de sentimento expressos em redes sociais. Fortuny et al. (2012) debruçaram-se sobre as questões políticas na Bélgica, num período de crise no final do ano de 2011, com o objetivo de medir o "sentimento social" dos cidadãos em relação aos partidos políticos do país, tendo por base os dados extraídos das notícias de grandes jornais em versões *online*.

Em Portugal, o trabalho realizado por Varela (2012) procura obter um sistema de classificação que seja independente da linguagem e realizou análises dos resultados através de bases de dados sobre cinema na língua espanhola e portuguesa (*Internet Movie Database* - IMDb). As experiências realizadas concluíram que o classificador Multinomial *Naive Bayes* é o que melhor se adequa à classificação de documentos mais pequenos.

Arunachalam & Sarkar (2013) propuseram um modelo e um estudo do caso para monitorizar e analisar o sentimento dos cidadãos nas redes sociais (Twitter, Facebook, YouTube, etc.) acerca das principais organizações que administram os apoios sociais nos EUA. No processo, os decisores obtêm informações valiosas, que podem ser convertidas em indicadores de gestão.

A utilização das redes sociais pelas administrações tributárias foi abordada por Dellia & Tjahyanto (2017), onde é estudada a utilização do Twitter para analisar as queixas / reclamações tributárias. O estudo revelou as dificuldades no tratamento automático de reclamações fiscais derivado da enorme quantidade de publicações não respeitantes à matéria em análise.

4. METODOLOGIA

A fase de extração dos *tweets* decorreu entre agosto de 2018 e fevereiro de 2019 (7 meses). Ultrapassada a fase de identificação do problema e de recolha dos dados, passámos à fase de pré-processamento, executando diversas tarefas de tratamento e padronização, seleção de termos relevantes e posterior transformação em formato estruturado para facilitação da extração de padrões. O reflexo da abrangência do tema em estudo na quantidade de *tweets* recolhidos, a par das limitações nele impostas quanto ao seu alcance, implicou o recurso a diversas técnicas de *text mining* de modo a restringir o universo dos *tweets* associados ao tema em análise. Efetuámos, também, a classificação manual de uma amostra aleatória de 1.015 casos, utilizada nos testes dos modelos selecionados sobre a qual foram efetuados os testes.

4.1. Extração de tweets

Na extração dos *tweets* utilizámos a *Application Programming Interface* (API) de pesquisa disponibilizada pela rede Twitter, recorrendo a um conjunto de instruções em linguagem R[1] para extração dos *tweets* em que estivessem presentes os termos (ou conjugações dos termos) "Autoridade Tributaria", "Autoridades Fiscais", "Finanças", "Fisco", "IRS", "IVA" ou "IRC". Os termos escolhidos para a pesquisa são aqueles que, na opinião dos autores, melhor se identificam com o problema e que têm maior probabilidade de estarem incluídos em comentários acerca do tema em análise.

[1] Linguagem de programação para gráficos e cálculos estatísticos (https://www.rstudio.com/)

A limitação temporal imposta pela rede social Twitter que impossibilita a extração de *tweets* com uma data de criação anterior a 10 dias em relação à data de extração, obrigou a execução periódica do conjunto de instruções "R" acima referido, exportando os resultados de cada extração para um ficheiro do tipo "CSV" (*Comma-Separated Values)*. No final do período programado para a extração de *tweets*, juntámos os registos de todos os ficheiros do tipo "csv" num único ficheiro, que designámos por "teste1.csv" e que contém o conjunto de dados para análise.

Utilizámos também o *software* RStudio[2], versão 3.5.2, na exploração do conteúdo dos *tweets*, por se tratar de um *software* gratuito e que contém vários recursos disponíveis para análise de textos. O *software* R carrega diversas bibliotecas e funções básicas no seu arranque, necessárias para o seu funcionamento. No entanto, no repositório CRAN (Comprehensive R Archive Network)[3] está disponível uma vasta gama complementar de bibliotecas (pacotes) com aplicação em projetos de *text mining*.

O pacote 'tm' do "R" possui as principais funções utilizadas no *text mining*, sendo o mais utilizado no tratamento e análise de textos. Trata-se, de facto, de uma estrutura para aplicações em *text mining* no "R" que inclui diversas funções pré-definidas (ex.: Corpus, tm_map, stopwords, termFreq, etc). O dicionário "readxl" é indispensável para a importação de ficheiros ".csv". Para além destas, utilizámos também outras bibliotecas cujas funções se explicitarão à medida da sua implementação.

4.2. Técnicas auxiliares para pré-tratamento

Durante a análise do ficheiro criado na etapa anterior (teste1 = twt1), verificámos que este continha 84.664 linhas correspondentes a outros tantos *tweets* inseridos na rede social Twitter entre 2018-07-29 e 2019-02-25, dos quais retirámos apenas os campos com informação pertinente:

- doc_id, correspondente ao número interno de identificação do *tweet*;
- texto, correspondente ao texto inserido pelo utilizador;
- created, correspondente à data de inserção do *tweet* na rede social;
- screenName, correspondente ao nome interno do utilizador do Twitter;
- retweetCount, correspondente à contagem de *retweets* do *tweet*.

Através de um conjunto de instruções destinadas a eliminar os registos não relacionados com a temática em estudo reunimos um conjunto de registos selecionados para a fase de pré-processamento corresponde a cerca de 30% da quantidade inicial de *tweets*. Esta tarefa foi realizada com recurso às

[2] RStudio é um software livre de ambiente de desenvolvimento integrado para R

[3] Rede de servidores ftp e web em todo o mundo que armazena versões idênticas e atualizadas de código e documentação para R

técnicas de contagem simples de termos mais frequentes. Efetuámos também algumas correções de erros ortográficos no texto em palavras com elevada frequência.

O passo seguinte consistiu na criação de um script em "R" que executa as transformações necessárias à criação do "corpus", agora num formato compatível para aplicação de técnicas baseadas na análise "*n-grams*"[4]. Nesta fase, e depois de agrupados os *tweets* repetidos num único registo (texto) que integrarão o "*corpus*"[5], são removidas as *stopwords*[6], a pontuação e os espaços em branco extra existentes no "*corpus*", para além da transformação das palavras em minúsculas. O conjunto de *tweets* (sem repetições) que correspondem ao "*corpus*" é composto por 12.663 registos

Apoiados nas técnicas que suportam a análise *n-grams*, listámos 300 sequências de duas palavras que ocorrem simultaneamente. O *output* gerado possibilita a eliminação de *tweets* que incluem "*bigrams*" que indiciam não serem relacionados com o objetivo deste estudo. Ao conjunto de dados resultante aplicámos o mesmo algoritmo, fazendo agora corresponder n=3. Em seguida, aplicámos a mesma técnica, fazendo corresponder n=2 e calculando a medida "TD-IDF" (Salton & Buckley, 1988) para os pares de termos. Estes procedimentos permitiram a exclusão de 650 *tweets* não relacionados com a temática em estudo. No final desta etapa obtivemos um conjunto de dados com 12.013 registos correspondentes a outros tantos *tweets* (sem repetições).

Recorrendo às técnicas *Part-of-Speech Tagging* ("POS-Tagging"), procedemos à exclusão de frases que não estão escritas na língua portuguesa. Para além das bibliotecas anteriormente descritas, estas técnicas são apoiadas em funções disponíveis nas bibliotecas "ptstem", "udpipe" e "textcat".

POS-Tagging é o processo de definição da categoria linguística das palavras (substantivos, verbos, adjetivos, etc.), através do seu comportamento morfossintático. Como o texto é uma fonte não estruturada de informação, para torná-lo como uma entrada adequada para um método automático de extração de informação é necessário transformá-lo num formato estruturado (Patheja, Waoo, & Garg, 2012).

Com a aplicação de um conjunto de instruções "R", carregámos o modelo de anotação de texto para a língua portuguesa ("portuguese-bosque-ud-2.3-181115.udpipe", disponível através da biblioteca "udpipe") e aplicámo-lo ao conjunto de dados obtido na etapa anterior, tendo sido eliminados 554 tweets que o algoritmo identificou como não redigidos na língua portuguesa e que comprovámos por observação direta do respetivo "output". O conjunto de dados assim gerado inclui 11.459 registos, com todas as palavras incluídas nos *tweets* anotadas de acordo com a sua categoria linguística.

---

[4] Sequência contígua de n itens de uma dada amostra de texto

[5] Coleção de documentos que contêm textos escritos em linguagem natural

[6] Palavras com alta frequência em textos e que normalmente não acrescentam conteúdo, tais como preposições, artigos, conjunções e outros.

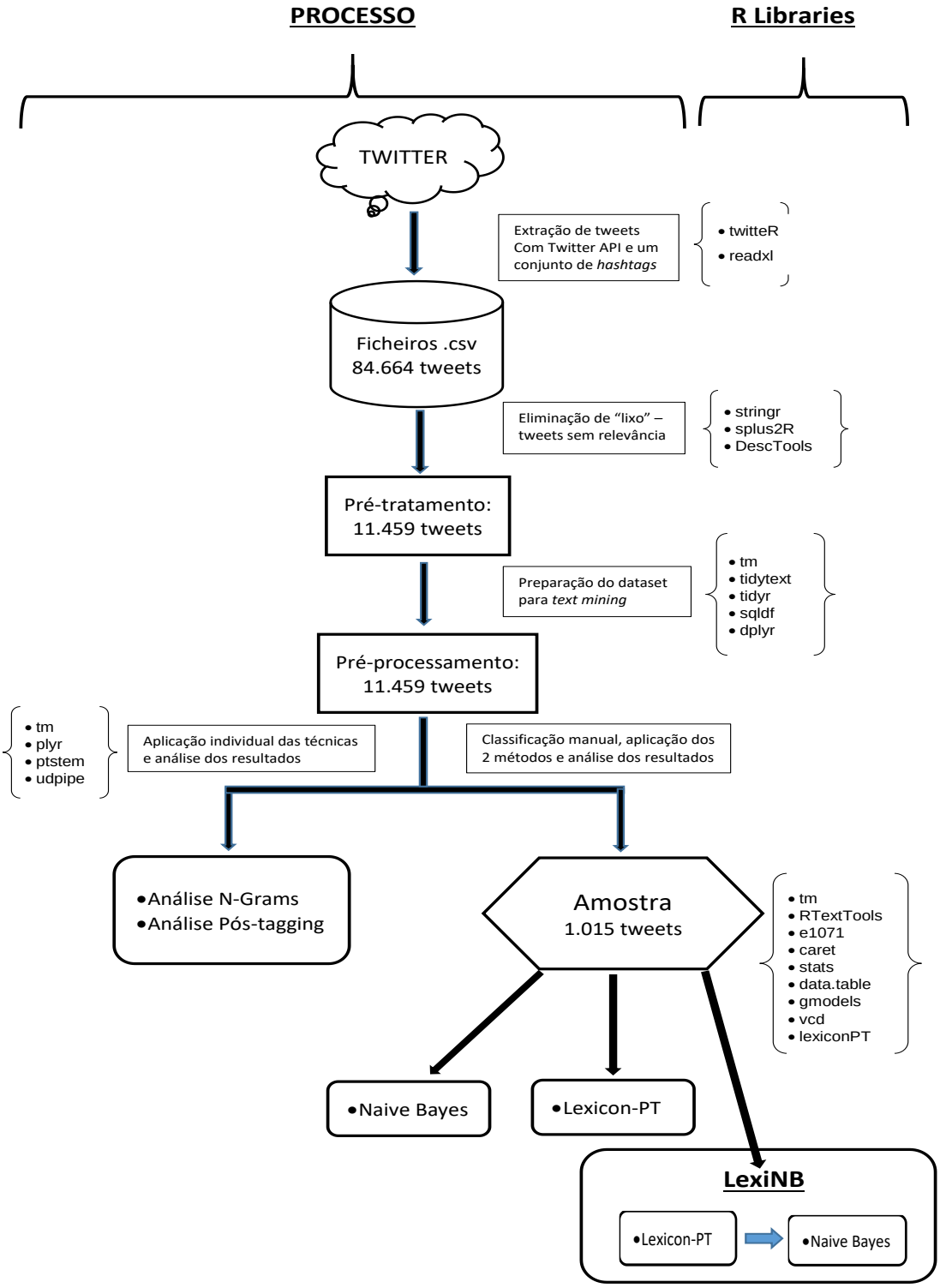

Figura 1 – Resumo da metodologia utilizada

A Figura 1 resume o processo de extração de *tweets*, pré-preparação e pré-processamento, para criação do conjunto de dados a utilizar, quer na análise *N-Grams* e *POS-Tagging* (secção 5), quer na extração da amostra utilizada nos testes com utilização do algoritmo "*Naive Bayes*" e da biblioteca "R" "Lexicon-PT".

4.3. Extração da amostra e classificação manual de tweets

A classificação de documentos não é um processo simples especialmente no caso dos *tweets* em que as mensagens são expressas através de textos muito curtos, com inúmeros erros ortográficos e em que muito poucos casos expressam uma opinião de forma clara. A fase inicial da classificação passa pela identificação e separação de textos do tipo informativo versus opinativo. Estes últimos são alvo de tratamento posterior quanto à sua objetividade, procurando-se identificar algumas componentes ou características que favoreçam o desempenho dos algoritmos de classificação

O processo de classificação manual dos *tweets* incidiu sobre uma amostra aleatória de 1.015 casos, gerada a partir do conjunto de dados criado na fase de pré-processamento. Exportada a amostra para um ficheiro Excel, nele realizámos a classificação manual dos *tweets* distinguindo aqueles que são relacionados com o tema em análise e simultaneamente manifestam uma opinião positiva (1) e negativa (-1), e os restantes que, por não respeitarem ao tema ou apenas descreveram um facto ou informação, são classificados com uma pontuação igual a 0. Tendo em conta a diversidade e o grau de subjetividade observada na expressão das opiniões, optou-se por considerar como comentário de teor positivo, todo aquele que expressasse uma crítica negativa acerca de um comportamento negativo relacionado com o tema (ex.: comentário negativo acerca de um indivíduo que praticou fraude fiscal).

Dos 1.015 registos que continha a amostra, apenas 353 expressavam uma opinião relacionada com o tema em análise, e destes, apenas 125 manifestavam opinião positiva (ou não negativa) representando cerca de 12% do total da amostra. Aos restantes 662 casos foi atribuída uma classificação neutra (class = 0),

5. EXTRAÇÃO DE CARACTERÍSTICAS

O recurso às análises *n-grams* e *POS-Tagging* em trabalhos desenvolvidos durante a fase de pré-processamento permitiu excluir um conjunto de 1.204 *tweets* não relacionados com a temática em estudo. O processo de extração de características foi direcionado no sentido de identificar os termos associados às palavras-chave utilizadas na pesquisa, tendo em vista a obtenção das características mais discriminantes para a classificação a efetuar.

5.1. Análise n-grams

A abordagem *n-grams* permite capturar a estrutura da linguagem do ponto de vista estatístico. A técnica consiste na criação de um "corpus" a partir dos *tweets* selecionados, mapeado e sob a forma de uma tabela de dupla entrada. Tendo em vista a análise evolutiva dos termos mais frequentes, efetuámos uma repartição do conjunto de dados inicial (84.664 registos) em 3 conjuntos:

- dt201809 – inclui os *tweets* cuja data de criação ocorreu entre 2018-07-29 e 2018-09-30;
- dt201812 - inclui os tweets cuja data de criação ocorreu entre 2018-10-01 e 2018-12-31;

- dt201902 - inclui os tweets cuja data de criação ocorreu entre 2019-01-01 e 2019-02-25.

Na figura 2 apresentamos os resultados para os termos mais frequentes em cada um dos períodos:

	Unigram201809	n	perct	Unigram201812	n	perct	Unigram201902	n	perct
1	iva	4152	0.0574	iva	4891	0.0539	iva	4461	0.0567
2	é	1715	0.0237	é	2290	0.0252	é	1934	0.0246
3	fisco	1383	0.0191	fisco	2058	0.0227	fisco	1607	0.0204
4	irs	1357	0.0188	irs	1687	0.0186	irs	1425	0.0181
5	touradas	825	0.0114	touradas	862	0.00950	touradas	842	0.0107
6	6	553	0.00764	6	587	0.00647	6	563	0.00716
7	vai	441	0.00609	irc	542	0.00598	vai	467	0.00594
8	irc	407	0.00562	vai	520	0.00573	irc	455	0.00579
9	sobre	383	0.00529	sobre	459	0.00506	sobre	406	0.00516
10	ser	342	0.00473	ser	433	0.00477	ser	370	0.00471
11	redução	311	0.00430	pagar	358	0.00395	pagar	316	0.00402
12	ps	304	0.00420	redução	335	0.00369	redução	316	0.00402
13	baixar	300	0.00415	baixar	323	0.00356	ps	306	0.00389
14	pagar	282	0.00390	ter	320	0.00353	baixar	304	0.00387
15	eletricidade	274	0.00379	ps	308	0.00340	eletricidade	275	0.00350
16	governo	263	0.00363	23	300	0.00331	23	274	0.00348
17	23	249	0.00344	governo	296	0.00326	ter	274	0.00348
18	ter	246	0.00340	eletricidade	283	0.00312	governo	270	0.00343
19	taxa	242	0.00334	taxa	276	0.00304	taxa	258	0.00328

Figura 2 – Análise evolutiva dos termos mais frequentes (output "R")

Com a utilização de *bigrams* (2 termos adjacentes) também não se registam alterações significativas nos termos mais frequentes para cada um dos diferentes períodos considerados na análise. De facto, o período temporal de extração de *tweets* é muito curto (7 meses), em que os temas dominantes estão relacionados com propostas para o Orçamento de Estado de 2019 para redução da taxa de IVA a aplicar nos ingressos em espetáculos tauromáquicos e nos consumos de eletricidade, para além da redução das taxas de IRS a aplicar a ex-emigrantes.

5.2. Análise POS-Tagging

Para Loper & Bird (2002), o objetivo da etiquetagem morfossintática (POS-Tagging) é de classificar os *tokens* de uma frase (ou texto) de acordo com suas classes morfológicas (substantivo, verbo, adjetivo, etc.). Para além da biblioteca "udpipe", necessária para a realização das tarefas de tokenização, *stemming*[7], lematização[8] e *POS-Tagging* de forma automática, utilizámos a biblioteca "lattice" para visualização dos resultados desta análise. Na figura 3 apresentamos um resumo da distribuição morfossintática do texto.

[7] Os algoritmos de stemming funcionam cortando o fim ou o início da palavra, levando em conta uma lista de prefixos e sufixos comuns que podem ser encontrados numa palavra.

[8] processo de agrupar as formas flexionadas de uma palavra para que elas possam ser analisadas como um único item, identificado pelo lema da palavra ou pela forma de dicionário.

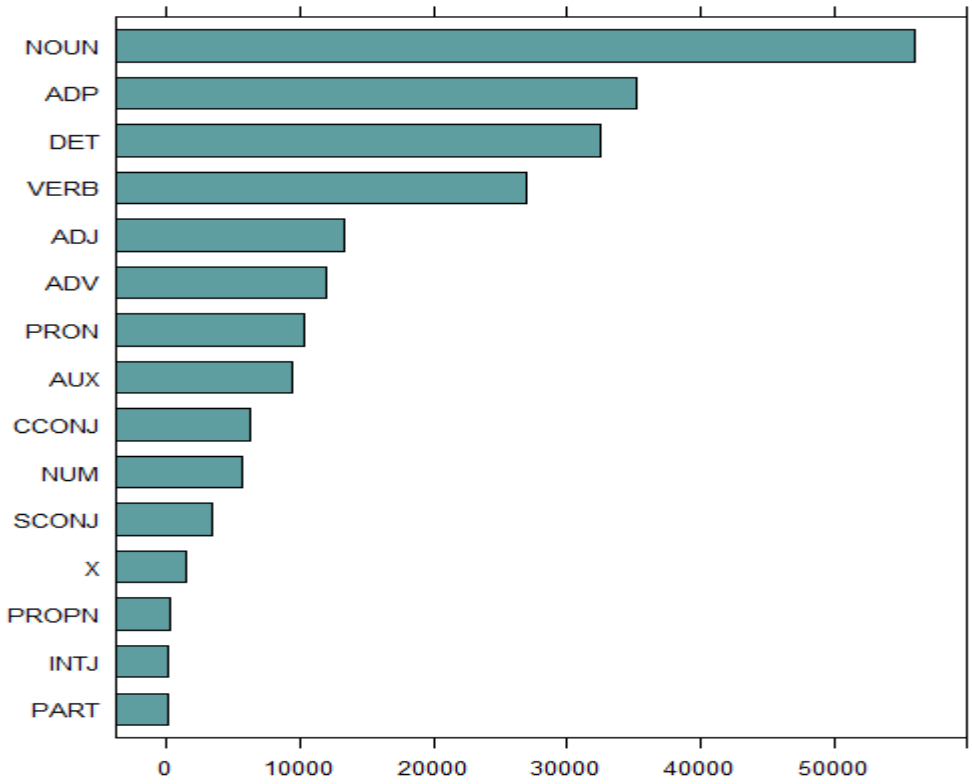

Figura 3 – Frequências das classes gramaticais

A quantidade de "nomes" presentes nos *tweets* selecionados representa cerca de 25% da totalidade dos *tokens* existentes, enquanto as preposições e determinantes têm um peso a rondar os 15% cada. Já os verbos representam apenas cerca de 12% da totalidade dos *tokens* e os adjetivos cerca de 6%.

A figura 4 apresenta os nomes com maior número de ocorrências. Nesta análise, optámos por retirar as *hashtags* incluídas na pesquisa inicial de *tweets*, devido à sua expectável relevância. Os resultados confirmam a tendência observada na análise *n-grams*.

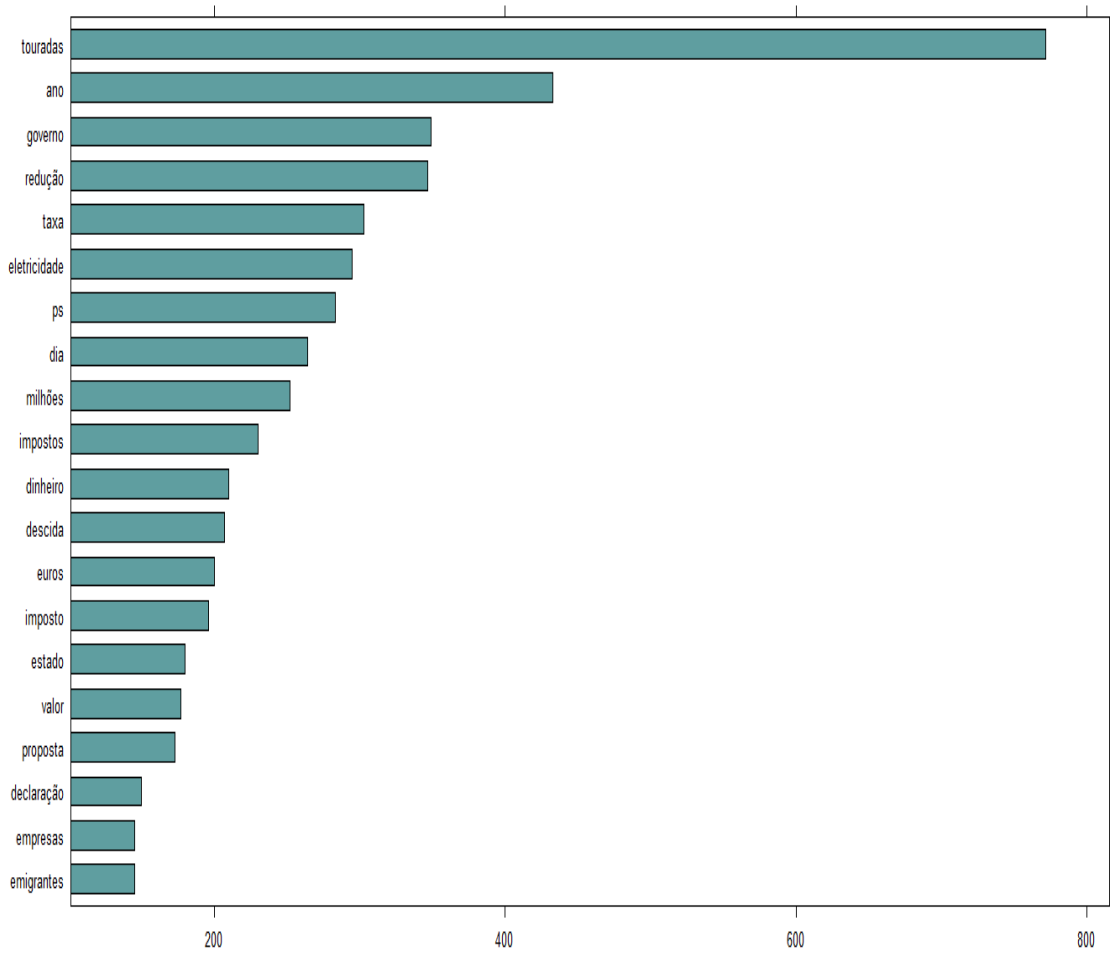

Figura 4 – Frequências de termos da classe gramatical “nomes”

Na análise efetuada ao conjunto de verbos com maior relevância no conjunto de registos, observámos a ocorrência de verbos diretamente relacionados com os impostos, como é o caso dos verbos “pagar” e baixar”. Efetuámos a mesma análise para a classe gramatical de adjetivos com maior número de ocorrências no conjunto de dados em análise. De realçar o facto de que o algoritmo não consegue distinguir a classe gramatical de alguns dos elementos do texto, como é o caso dos termos “fiscais”, “fisco”, “fiscal” e “tributária” que, em determinado contexto poderão ter a categoria de “adjetivo”, mas que, no tema em análise, terão a classe gramatical de “nome”.

Recorrendo às funções disponíveis nas bibliotecas “tm” e “igraph” e outras, procedemos à análise de coocorrências, com várias simulações para pares “nome – adjetivo”, “nome – verbo” e “verbo – adjetivo”. O processo implica a criação de um “corpus” mapeado, em que o grau de coocorrência é medido por uma função de verosimilhança (Log-likelihood). A representação gráfica desta medida reflete-se na dimensão dos círculos de cor amarelada. Na figura 5 apresentamos os resultados deste método para o termo “iva”.

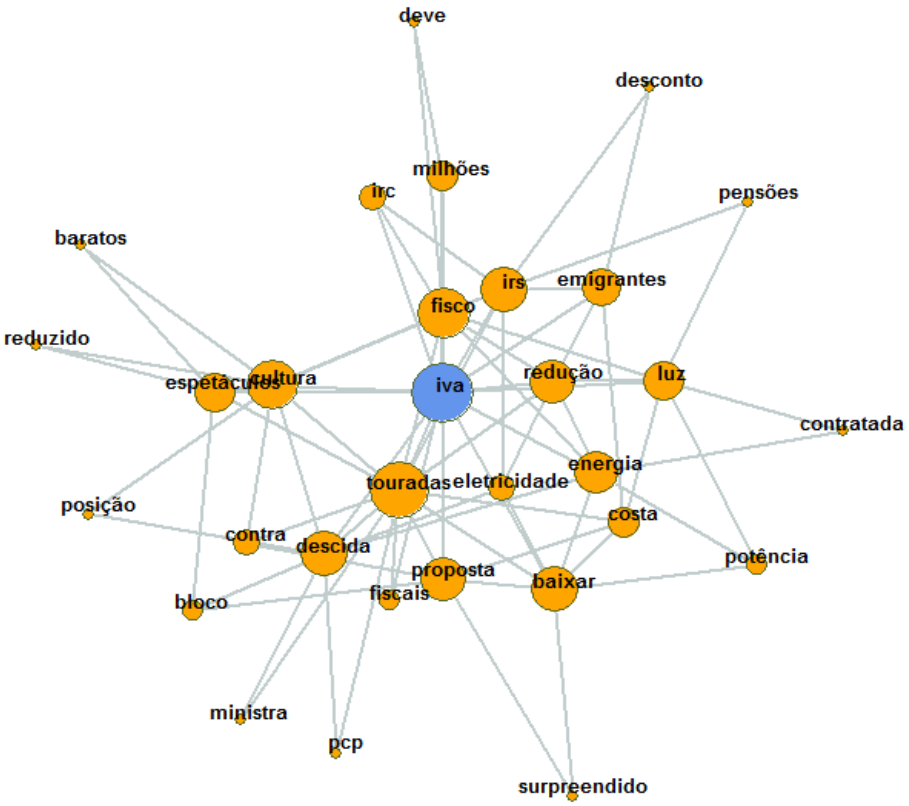

Figura 5 – Grafo de coocorrências para o termo "IVA"

O termo "iva" apresenta maiores índices de coocorrência com as palavras "touradas", "cultura", "descida" e "fisco", traduzindo o vasto conjunto de *tweets* publicados relacionados com a temática da redução do IVA aplicável em ingressos de espetáculos tauromáquicos.

Realizámos o mesmo procedimento para os termos "irs" e "fisco", excluindo as restantes *hashtags* utilizadas na pesquisa. Relativamente ao termo "irs", este apresenta maiores índices de coocorrência com as palavras "tabelas", "retenção", "2019", "salários", "pensões", "emigrantes" e "desconto", refletindo as preocupações manifestadas pelos utilizadores desta rede social em relação às retenções de IRS aplicáveis em 2019, mas também em relação ao novo regime de IRS aplicável a ex-emigrantes (artigo 12º-A do CIRS). Já o termo "fisco" apresenta maiores índices de coocorrência com as palavras relacionadas com dívidas fiscais ("deve", "dívida", "dívidas", milhões).

6. RESULTADOS

Para cumprimento do objetivo proposto neste trabalho, utilizámos o algoritmo *Naive Bayes* e a biblioteca "Lexicon-PT" do software "R", numa abordagem bietápica que designámos de "LexiNB". Como verificaremos mais adiante, a utilização combinada dos dois modelos gerou resultados mais satisfatórios do que os alcançados por qualquer um deles isoladamente.

6.1. Resultados para Naive Bayes utilizando a classificação manual

Neste trabalho, aplicámos o algoritmo *Naive Bayes* a uma amostra aleatória de 1.015 casos, gerada a partir do conjunto de dados criado na fase de pré-processamento e manualmente classificada. A realização desta operação envolve a transformação do corpus correspondente à referida amostra, num formato que possibilite a realização de análises quantitativas, como é o caso dos objetos do tipo "Document Term Matrix". Nestas matrizes, as linhas correspondem aos documentos na coleção de textos e as colunas correspondem aos termos. Existem várias medidas para o valor que cada entrada na matriz (ex: TD-IDF). A este objeto, aplicámos uma restrição aos termos com 2 ou mais ocorrências ("freq <- findFreqTerms(dtm.train, 2)").

De seguida, procedeu-se à repartição num conjunto de treino correspondente a 67% dos registos que compõem a amostra, e os restantes 33% dos registos foram incluídos no conjunto de teste. Recorrendo à função "naiveBayes" da biblioteca "R" "e1071", gerámos um classificador que permitiu prever a classificação dos *tweets* na amostra de teste.

Na aplicação do modelo utilizámos o suavizador de Laplace, de modo a contornar o problema da probabilidade condicionada quando os novos dados incluem valores de recurso que nunca ocorrem para um ou mais níveis de uma classe de resposta. O suavizador de Laplace adiciona um pequeno número a cada uma das contagens nas frequências de cada recurso, o que garante que cada recurso tenha uma probabilidade diferente de zero para cada classe. Normalmente, um valor de um ou dois para o suavizador de Laplace é suficiente.

Verifica-se que o significativo grau de assertividade total do modelo (78,87%), que designaremos por "**NBayes FTerms2**", só é alcançado à custa da previsão efetuada para *tweets* não relacionados com o tema ou sem opinião (class = 0) conforme atesta o indicador de sensitividade. Este facto é também confirmado pela elevada percentagem do indicador "No Information Rate" na escolha da classe maioritária. A estatística "Kappa", igual a 0.0245, indica uma fraca precisão do classificador utilizado em função da sua precisão esperada. Em todo o caso, o grau de assertividade total do modelo (78,87%) é superior à proporção de elementos incluídos na classe maioritária que representam 65,2% do total da amostra.

Aplicando uma restrição ao número de termos com 3 ou mais ocorrências ("freq <- findFreqTerms(dtm.train, 3)") e mantendo o conjunto de treino correspondente a 67% da amostra, reduzimos ligeiramente a assertividade do modelo (74,5%), que designaremos por "**NBayes FTerms3**", mas aumentamos a sensitividade e especificidade na classificação de *tweets* não neutros. Observa-se que também um decréscimo no valor do indicador "No Information Rate", revelador de uma ligeira melhoria, confirmada pelo acréscimo registado na estatística "Kappa", indicando um ligeiro aumento da precisão do classificador. De salientar também a melhoria da assertividade individual em todas as classes.

Nas figuras seguintes apresentamos os histogramas das distribuições das 3 classes utilizadas no modelo:

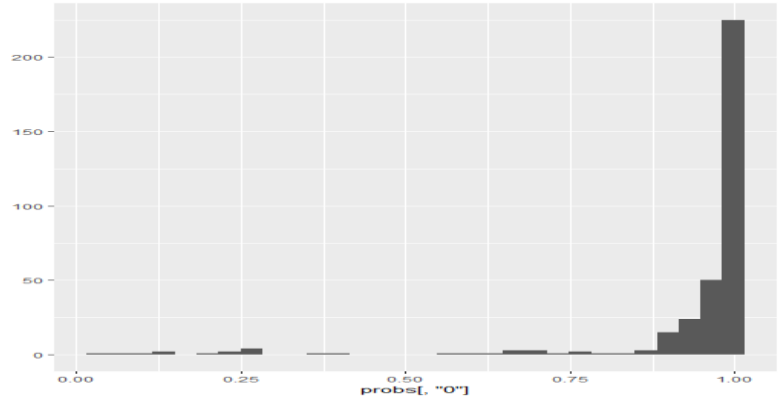

Figura 6.A – Histograma de distribuição da classe 0 (class = 0) do modelo

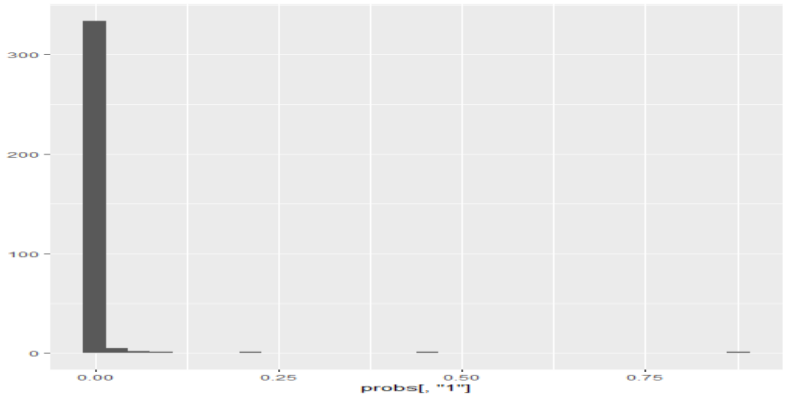

Figura 6.B – Histograma de distribuição da classe 1 (class = 1) do modelo

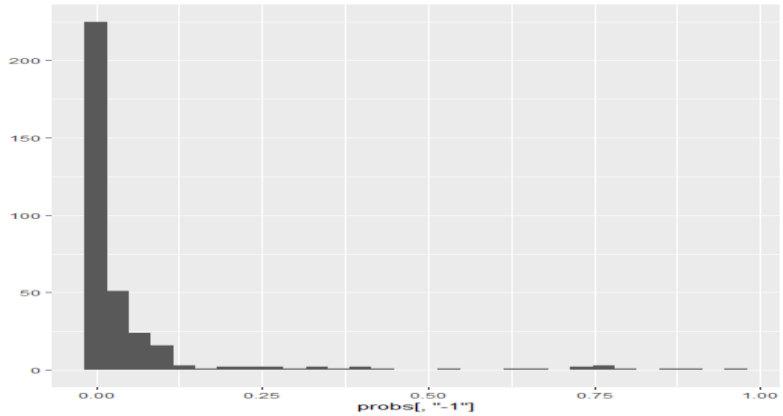

Figura 6.C – Histograma de distribuição da classe 1 (class = 1) do modelo

Nos *tweets* com classificação neutra (*class* = 0), a probabilidade aproxima-se do valor 1 à medida que as frequências vão aumentando, enquanto a classificação com opiniões negativas (*class* = -1) regista um comportamento inverso. Já os *tweets* com opiniões positivas (*class* = 1) concentram a grande maioria das suas frequências numa probabilidade igual a zero.

6.2. Resultados com utilização do Lexicon-PT

O "Lexicon-PT" é atualmente uma biblioteca disponível no CRAN e dispõe de dois léxicos polarizados: "oplexicon-PT" (Souza & Vieira, 2012) e "sentiLex-PT" (Carvalho & Silva, 2015).

Na aplicação dos léxicos incluídos na biblioteca "Lexicon-PT" à amostra aleatória de 1.015 casos, efetuámos uma simulação com atribuição de polaridade apenas em palavras que constassem dos 2 léxicos (oplexicon e sentiLex-PT) em simultâneo e um outro teste em que atribuímos a polaridade aos termos que constassem num dos 2 léxicos.

Na execução do primeiro teste apenas foram classificados 267 *tweets*, permanecendo 748 por classificar neste modelo que designámos por "**LexPT Oplex**". Pela análise tabular dos *tweets* classificados, verificámos que o modelo atribuiu pontuações inteiras variáveis entre -3 e 2. Tendo em vista a harmonização da escala comparativa entre modelos, optámos por estabelecer uma regra do tipo:

- Se (Classificação_Inicial > 1), então Classificação_Final =1
- Se (Classificação_Inicial < -1), então Classificação_Final =-1
- Noutros casos, então Classificação_Final = 0

Designámos de "Lexicon-PT Oplexicon" o modelo gerado com base na classificação atribuída pelo léxico "oplexicon", e o modelo gerado com base na classificação atribuída pelo léxico "sentiLex-PT" foi designádo de "**LexPT sentLex**".

Os resultados obtidos nestes dois testes são inferiores, em toda a linha, aos obtidos com o algoritmo Naive Bayes, já que classificaram apenas 267 dos 1.015 casos da amostra (26,3%).

Para a realização de um terceiro teste recorremos à biblioteca "Lexicon-PT" e utilizámos a mesma amostra, com polaridade atribuída por qualquer um dos léxicos utilizados, aos termos presentes na amostra. A estratégia consistiu na criação de duas tabelas auxiliares, com instruções do tipo "*inner join*", ligando os termos que constam da amostra aos termos presentes em cada um dos léxicos utilizados. A junção das duas tabelas gerou alguns elementos sem pontuação atribuída ("NA") numa das variáveis e que foram substituídas pela classificação atribuída pelo outro léxico. Também neste caso se verificou que o modelo atribuiu pontuações inteiras variáveis entre -3 e 2, tendo-se aplicado a regra de classificação igual à utilizada no modelo anterior.

A pesquisa originou um conjunto de 880 *tweets* classificados neste modelo, que designámos de "**LexPT SentOp**", sendo que: 52 deles foram classificados positivamente, igual quantidade foi classificada de forma negativa e 776 foram classificados como neutros (sent = 0). Nota-se, aqui, uma ligeira melhoria dos resultados em relação ao teste anterior. Contudo, os resultados obtidos permanecem inferiores aos obtidos com o algoritmo *Naive Bayes*. O número de casos classificados neste modelo aumentou significativamente para 86,7% do total da amostra.

Experimentámos também a utilização do algoritmo *Naive Bayes*, aplicado ao *conjunto de dados* gerado na etapa anterior (880 *tweets*). O método utilizado consistiu na transformação desse conjunto de dados num objecto do tipo "Document Term Matrix" restringido aos termos com 3 ou mais ocorrências ("freq <- findFreqTerms(dtm.train, 3)") e posteriormente repartido num conjunto de treino correspondente a 66% dos registos e os restantes 34% foram incluídos no conjunto de teste. A este modelo atribuímos a designação de "**LexiNB**".

6.3. Resumo comparativo dos resultados

Na tabela 2 apresentamos um resumo dos principais indicadores estatísticos obtidos na aplicação dos 6 modelos testados:

Indicador	**NBayes FTerms2**	**NBayes FTerms3**	**LexPT Oplex**	**LexPT sentLex**	**LexPT SentOp**	**LexiNB**
Accuracy	0,7420	0,7449	0,6479	0,6367	0,6841	0,7483
No Information Rate	0,7391	0,7391	0,6966	0,6966	0,7625	0,7483
Kappa	0,0236	0,1039	0,0160	-0,0077	-0,0074	0,2897
Sensitivity						
Class: -1	0,0179	0,0893	0,1053	0,0877	0,0638	0,4902
Class: 0	1,0000	0,9804	0,8978	0,8871	0,8763	0,8879
Class: 1	0,0000	0,0588	0,0000	0,0000	0,0735	0,0000
Specificity						
Class: -1	0,9965	0,9758	0,9571	0,9571	0,9418	0,8866
Class: 0	0,0222	0,1000	0,0988	0,0864	0,1005	0,3733
Class: 1	1,0000	1,0000	0,9506	0,9424	0,9421	1,0000
Detection Rate						
Class: -1	0,0029	0,0145	0,0225	0,0187	0,0102	0,0839
Class: 0	0,7391	0,7246	0,6255	0,6180	0,6682	0,6644
Class: 1	0,0000	0,0058	0,0000	0,0000	0,0057	0,0000
Balanced Accuracy						
Class: -1	0,5072	0,5325	0,5312	0,5224	0,5028	0,6884
Class: 0	0,5111	0,5402	0,4983	0,4868	0,4884	0,6306
Class: 1	0,5000	0,5294	0,4753	0,4712	0,5078	0,5000

Tabela 2 – Resumo comparativo dos resultados

Comparado com s resultados obtidos com o melhor dos modelos ("NBayes FTerms3"). o modelo designado de "LexiNB" regista um aumento significativo (179%) da estatística *Kappa* que mede a precisão do classificador utilizado em função da sua precisão esperada, situando-se agora num escalão moderado, e um incremento médio de cerca de 14% na sua eficiência (Balanced Accuracy = (sensibilidade + especificidade) / 2), que resulta essencialmente do aumento na deteção de opiniões

negativas, permanecendo, no entanto, com indicadores muito baixos. Observa-se também, uma ligeira melhoria na assertividade total, quando comparada com a utilização isolada dos modelos *Naive Bayes* ou do dicionário Lexicon-PT.

7. CONCLUSÕES

A limitação do número de carateres posta pela rede social Twitter introduz algumas limitações na análise "*n-grams*" impossibilitando a geração de *n-grams* com n superior a 4, uma vez que a sua quantidade é muito diminuta. Também as técnicas de "POS-Tagging" estão, de alguma forma, comprometidas no nosso trabalho derivado da grande desproporção entre a quantidade de "nomes" e de "verbos" ou "adjetivos". A elevada quantidade de erros ortográficos e a inexistência de corretores eficazes para a língua portuguesa, para além da pequena quantidade de *tweets* que expressam opinião de forma clara, são outras das limitações que encontramos no decurso deste trabalho.

Na nossa análise observámos que os resultados obtidos individualmente, quer pelo algoritmo *Naive Bayes* quer pela utilização da biblioteca "Lexicon-PT" não são suficientemente interessantes do ponto de vista estatístico. Realçamos, contudo, a melhoria significativa dos resultados obtidos com a abordagem bietápica proposta neste artigo, designada de "LexiNB", em que aplicámos o algoritmo *Naive Bayes* ao conjunto de dados gerado com a utilização dos dois léxicos disponíveis na biblioteca "Lexicon-PT".

Este estudo permite concluir que os utilizadores da rede social Twitter abordam, com maior predominância, a temática dos impostos e da carga fiscal, sendo menos frequentes os comentários sobre o funcionamento da administração fiscal portuguesa, sobressaindo, nestes últimos, a preocupação dos utilizadores em relação aos elevados montantes das dívidas ao fisco. Da análise dos resultados suportados nas previsões obtidas com recurso à abordagem bietápica "LexiNB", concluímos que as opiniões negativas se manifestam essencialmente acerca da redução da taxa de IVA nas touradas, enquanto as opiniões positivas se manifestam através de críticas negativas acerca de comportamentos negativos relacionados com a fraude em IVA e com dívidas ao fisco.

Em trabalhos futuros, perspetivamos um alargamento do tamanho da amostra manualmente classificada e a aplicação do modelo misto atrás descrito como forma de colmatar as insuficiências encontradas e de melhorar o desempenho global na classificação dos *tweets*. Também não está posta de parte a hipótese de utilização de algoritmos de aprendizagem não supervisionada que possam contribuir para a melhoria do desempenho do nosso modelo.

AGRADECIMENTOS

Este trabalho foi apoiado pela FCT - Fundação para a Ciência e Tecnologia no âmbito do Projeto UID/CEC/00319/2019.

REFERÊNCIAS

Arunachalam, R., & Sarkar, S. (2013). The New Eye of Government: Citizen Sentiment Analysis in Social Media. Workshop on Natural Language Processing for Social Media (SocialNLP) (pp. 23-28). Nagoya, Japan,: IJCNLP.

Carvalho, P., & Silva, M. J. (2015). Sentilex-PT: Principais características e potencialidades. Linguística, Informática e Tradução: Mundos que se cruzam, Oslo Studies in Language 7(1), pp. 425-438.

Dellia, P., & Tjahyanto, A. (2017). Tax Complaints Classification on Twitter Using Text Mining. IPTEK, Journal of Science, Vol. 2, No. 1.

Filho, J. A. (2014). Mineração de textos: análise de sentimento utilizando tewwets referentes à copa do mundo 2014. Qixada, Ceará, Brasil: Universidade Federal do Ceará.

Fortuny, E. J., Smedt, T. D., Martens, D., & Daelemans, W. (2012). Media coverage in times of political crisis: A text mining approach. ELSEVIER - Expert Systems with Applications - www.elsevier.com/locate/eswa, 11616–11622.

Forum on Tax Administration. (2011). Social Media Technologies and Tax Administration. OECD - Centre for Tax Policy and Administration.

Hu, M., & Liu, B. (2004). Mining and Summarizing Customer Reviews. Department of Computer Science - University of Illinois at Chicago.

Indurkhya, N., & Damerau, F. J. (2010). Handbook of Natural Language Processing - Second Edition. USA: Chapman & Hall/CRC.

Liu, B., & Zhang, L. (2012). A Survey of Opinion Mining and Sentiment Analysis. Em B. Liu, & L. Zhang, Mining Text Data. USA.

Loper, E., & Bird, S. (2002). NLTK: The Natural Language Toolkit. Pennsylvania, Philadelphia, USA: Department of Computer and Information Science University of Pennsylvania.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classi¯cation using Machine Learning Techniques. Conference on Empirical Methods in Natural Language Processing (pp. 79-86). Philadelphia, USA: Association for Computational Linguistics.

Patheja, P., Waoo, A., & Garg, R. (2012). Analysis of part of speech tagging. International Conference on Intuitive Systems & Solutions (ICISS). Mumbai, INDIA: International Journal of Computer Applications® (IJCA).

Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the Poor Assumptions of Naive Bayes Text Classi. Twentieth International Conference on Machine Learning. Washington DC: Artificial Intelligence Laboratory - MIT - Cambridge.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, pp. 513-523.

Souza, M., & Vieira, R. (2012). Sentiment Analysis on Twitter Data for Portuguese Language. https://www.researchgate.net/publication/262175717.

Varela, P. d. (2012). Sentiment Analysis. Lisboa: Instituto Superior Técnico.