Análise de Sentimentos em comentários na língua portuguesa: uma comparação de métodos

Graziella C. Bonadia*, Gilmar Barreto**

*Fundação CPqD (e-mail: bonadia@cpqd.com.br).

** Faculdade de Engenharia Elétrica e de Computação – FEEC, Universidade Estadual de Campinas, (e-mail: gbarreto@dsif.fee.unicamp.br).

Abstract: This article presents a study on the sentiment analysis applied to films reviews dataset and news comments dataset in Portuguese language, based on the comparison of performance indicators for an English language dataset. Two widely used feature engineering methods, the bag of words (BoW) and the word embedding (MWE), were applied. The results of the experiments show positive and negative aspects of both approaches and emphasize the power of word density in the performance indicator, both for the BoW method and for the word dictionary formulation of word embedding.

Resumo: Este artigo apresenta um estudo sobre análise de sentimentos aplicada a uma base de comentários de filmes e de notícias em língua portuguesa, tendo como base de comparação os indicadores resultantes para uma base de língua inglesa. Foram aplicados dois métodos de engenharia de atributos bastante difundidos, o *bag of words* (BoW) e o *word embedding* (MWE). Os resultados dos experimentos mostram pontos positivos e negativos das duas abordagens e evidencia o poder da densidade de palavras nos resultados de acurácia, tanto para o método BoW quanto para a formulação do dicionário de palavras do *word embedding*.

Keywords: Sentiment analysis; word embedding, natural language; reviews.

Palavras-chaves: Análise de sentimentos; vetores de palavras; linguagem natural; comentários.

1. INTRODUÇÃO

Na última década, a capacidade de processamento e armazenamento das máquinas aumentou muito, assim como a popularização dos meios de comunicação em massa, que hoje independe dos canais de rádio e TV. Os provedores de conteúdo estão dispersos no meio da população, que expressa suas opiniões em comentários na internet, além de produzir notícias e entretenimento. Toda essa evolução da tecnologia demanda novas necessidades e que, atendidas, geram novas demandas, perpetuando um ciclo infundável de criação e resolução de problemas. Atualmente, com as técnicas de processamento de linguagem natural (PLN, ou, do inglês, NLP, *Natural Language Processing*), já é possível prever sentimentos sem necessitar um olhar humano em textos, áudio ou vídeos. As pessoas hoje em dia tendem a se expressar muito por meio de redes sociais e empresas como Google e Facebook, por exemplo, já utilizam o conteúdo produzido por seus usuários para criar modelos preditivos poderosos quanto a anúncios que devem aparecer individualmente, por exemplo.

Os textos disponíveis em diversos meios é uma fonte muito rica de informação para explorar o uso de tecnologias de automação. Há uma variedade de aplicações que podem ser desenvolvidas a partir do reconhecimento do conteúdo de textos e áudios em linguagem natural que devem promover cada vez mais a velocidade com que um determinado

objetivo será alcançado. Um exemplo bastante antigo e ilustrativo é a classificação de mails como legítimos ou *spams*. A partir da análise de palavras e expressões, o algoritmo *anti-spam* classifica uma mensagem e direciona o e-mail à pasta adequada. O processamento de linguagem natural também é utilizado para tradução automática, recuperação de informações (mecanismos de buscas) e inferências a partir de textos como, por exemplo, análise de intenção ou sentimentos.

Uma das aplicações que estão sendo estudadas no momento é a análise de sentimento de textos curtos como comentários de postagens nas redes sociais. Para este último, Tang *et al.* (2009) pode-se obter uma visão geral sobre as questões que permeiam a detecção de sentimentos em avaliações de maneira geral. Hoje, essa forma de socialização é muito utilizada pelo público que provê suas opiniões sobre um determinado *post*. Essas opiniões são utilizadas para melhorar a forma de apresentação de um conteúdo, direcionar assuntos de futuras postagens, avaliar popularidade ou aceitação de políticas, direcionamento de propagandas, previsão ações de terror, dentre inúmeras outras utilidades. Para avaliação de sentimentos desses textos curtos já podem ser encontrados diversos estudos em diferentes línguas como, por exemplo, em Rotim e Šnajder (2017).

Independentemente da língua em que o estudo é conduzido, existem inúmeras questões que impõem dificuldade no momento de criar uma máquina que automatize a predição de

sentimentos. A principal delas é a subjetividade da linguagem natural, que, além de ser dinâmica, inclui nuances como gírias, ironias, informalidades na escrita e fala, podendo a levar confusões até mesmo sob o olhar humano. Em Hercig e Lenc (2017) é discutido o impacto que a linguagem figurada tem na análise de sentimentos e apresenta uma comparação de resultados considerando métodos que utilizam e não utilizam informação adicional sobre sarcasmo. Já em Ding *et al.* (2008) é proposta uma abordagem que inclui palavras que dependem do contexto. E é nesse ponto que abordagens que incluem uma noção de contexto ou de significado do léxico podem trazer ganhos na acurácia de máquinas automatizadas de avaliação de sentimentos.

Um dos campos de estudo nessa linha foi trazido por Bengio (2003) com o conceito de *word embedding*, ou seja, uma forma de transformar palavras em vetores de dimensão parametrizável, sendo que seu resultado depende do contexto “médio” em que a palavra é utilizada. No entanto, ainda existia o alto custo de processamento para obter tais vetores. Assim foram surgindo novos métodos computacionais e estatísticos que provessessem os vetores n-dimensionais com um custo computacional mais viável (ver, por exemplo, Mikolov *et al.*, 2013). A partir de então, foram surgindo vários estudos versando sobre *word embedding* como Jin *et al.* (2016) e Yu *et al.* (2017), dentre inúmeros outros.

Todas essas discussões tem o objetivo de extrair as características principais que possam levar mais poder discriminante no momento de utilizar uma máquina qualquer de classificação. Embora essas máquinas genéricas de classificação devam ser utilizadas da melhor forma possível para potencializar os resultados, a principal contribuição para a acurácia dos resultados recai na escolha do vetor de atributos. Uma vez selecionados os atributos (*features*) a serem utilizados, torna-se praticamente automática a seleção do método de classificação, que pode contar com uma heurística de testes para otimização de modelos e hiperparâmetros.

Este artigo tem o objetivo de explorar as diferenças em acurácia de um classificador de sentimentos binário (positivo ou negativo) considerando variações nos métodos de criação de atributos e na relação entre as bases de treino e teste. Para isso, foi criada uma base de comentários sobre filmes na língua portuguesa do Brasil. Como forma de obter uma linha de base (*baseline*), todos os resultados serão comparados com o mesmo método realizado para uma base de comentários em língua inglesa muito difundida para uso em pesquisa (distribuída pelo site de competição do *Kaggle*).

Serão discutidas outras duas características sobre os métodos: a utilização de uma base de treino em inglês para classificação em português e a utilização de uma base de treino não relacionada com a base de teste. Para o treinamento, serão utilizados métodos de classificação com os hiperparâmetros pré-selecionados, não sendo, portanto, alvo de discussão neste artigo.

Esses dois últimos experimentos visam explorar os resultados nos casos em que não há dados suficientes para criar uma boa base de treinamento, então, como fonte da classificação de sentimentos, são usados modelos previamente treinados em outros contextos. No presente caso, um deles tem assunto semelhante, ou seja, avaliações de filmes, mas em língua inglesa e o outro, o assunto não tem nenhuma relação com a base utilizada para o modelo. São comentários provenientes de um site de notícias brasileiro.

2. PRÉ-PROCESSAMENTO E EXTRAÇÃO DE ATRIBUTOS

Geralmente, os primeiros passos de PLN incluem a obtenção de atributos com uma técnica chamada *tokenizing* (aqui, definido por tokenização). Em geral, essa é a primeira etapa do pré-processamento de textos. Esse processo é definido como “dividir o texto em unidades”, que podem ser palavras, números ou pontuações. Esse processo já permite a realização de análise de sentimentos, pois já possibilita o cálculo da distribuição de palavras em um texto. No entanto, apesar de a tokenização fornecer já uma forma mais interpretável pelas máquinas, ela provê informação limitada dependendo dos objetivos desejados do PLN. Em geral, existem outros métodos de pré-processamento que podem acompanhar a tokenização adicionando elementos mais informativos como o *tagging*, ou etiquetação, que consiste em associar à palavra (*token*) sua classe gramatical (definição segundo a morfologia) ou o *stemming*, que tem o objetivo de possibilitar o agrupamento de palavras que se diferenciam apenas pela inflexão. Palavras como “apresentou”, “apresentava” ou “apresentando” se tornam “apresent”. Há, também, o *lemmatizing*, que faz a troca de palavras por sinônimos para padronizar a linguagem. Esses procedimentos tendem a diminuir ruídos na massa de dados e podem melhorar a acurácia do objeto de estudo, mas também podem piorar.

Embora tais métodos possam ser utilizados para análise de sentimentos, o presente estudo apenas considerou a tokenização com remoção de *stopwords* (palavras muito comuns como “de”, “para”, “o”, “a”, etc.) para que estes não representem a maior parte das palavras no vetor de atributos. A lista de *stopwords* em inglês e em português utilizada foi obtida pela biblioteca NLTK.

Após a tokenização, segue a etapa de extração de atributos (ou características), que foi realizada de dois métodos distintos para comparação: o *Bag of Words* e o *Word Embedding*, com o método Word2Vec (Mikolov *et al.*, 2013). A próxima seção traz uma breve descrição de cada um desses métodos.

2.1 Contagem de palavras (Bag of Words)

Bag of words (contagem de palavras, em livre tradução) é um método bastante direto para a extração de atributos de um texto qualquer. Nele, há a contagem de cada palavra contida no texto formando um vetor em que cada elemento corresponde à frequência de uma palavra no texto. A Fig. 1 ilustra as etapas para a transformação do *corpus* em um vetor de frequência das palavras mais comuns.

Fig. 1 Construção dos atributos para o método *bag of words*.

Tomando, por exemplo, um *corpus* com avaliações sobre um livro (A). Cada avaliação, demonstrada em (A), é um parágrafo e passa pelo processo de tokenização com a retirada dos *stopwords*, fazendo com que as avaliações se transformem em um conjunto de palavras soltas (B). Cada palavra que aparece no *corpus* é associada a uma posição em um vetor (C), que é preenchida com a quantidade de vezes que aparece na avaliação (D).

Neste exemplo, cada avaliação é associada a um vetor no tamanho do número de palavras contidas em todo o *corpus*. No entanto, é comum limitar esse número para simplificação, já que há pouco poder discriminante em uma palavra que aparece poucas vezes ao longo do texto. Assim, no método *bag of words*, uma das configurações possíveis é a seleção do número de elementos deste vetor, ou seja, o número das palavras mais frequentes no *corpus*. Em resumo, os processos obrigatórios para classificação de avaliações utilizando *bag of words* são: criação de *tokens*, definição de tamanho do vetor e contagem de palavras.

Com essas ações, cada avaliação foi transformada em um vetor de contagem de palavras que servirá como variáveis de entrada em um modelo de classificação. A variável de saída (ou variáveis resposta) para a análise de sentimento pode ser uma nota ou simplesmente uma indicação de avaliação positiva, negativa, ou até mesmo neutra. Para a comparação de métodos deste artigo, foi utilizada a quantidade de 5.000 elementos.

2.2 Word Embedding

Embedding é um conceito matemático clássico e que foi emprestado ao tema relacionado ao processamento de linguagem natural. De maneira geral, o *embedding* é a conversão de um objeto em outro com a preservação de sua estrutura (f: $X \rightarrow Y$). No aprendizado de máquina, o *embedding* pode ser visto como um processo de redução de dimensionalidade com esta preservação de estrutura.

Em Bengio (2003), é tratado justamente esse ponto no processamento de linguagem natural. Uma palavra dentro de um *corpus* qualquer pode ser representada inicialmente por um elemento em um espaço vetorial. Essa palavra pode estar seguida de outra palavra e outra que a antecede. Em outra sentença do *corpus*, essa mesma palavra está seguida por uma palavra diferente da primeira, fazendo com que haja um número exorbitante de combinações possíveis.

A proposta de Bengio (2003), é transformar o espaço vetorial esparsos em um espaço denso com a diminuição da dimensionalidade por meio de uma rede neural recorrente. Imaginando, por exemplo, que o objetivo é prever a próxima palavra em uma frase. A probabilidade conjunta de ocorrer uma palavra w pode ser descrita como o produto das probabilidades condicionais de ocorrer a palavra w , dado todas as outras que ocorreram antes dela. Dessa forma, o número de combinações possíveis seria, além de extremamente alto, pouco generalizável, visto que a probabilidade de se obter uma sequência de palavras exatamente igual ao longo do *corpus* é mínima. Assim, Bengio propôs uma simplificação que utilizaria apenas a probabilidade condicional de uma palavra, dado um número pequeno de palavras imediatamente anteriores. Com isso, ele trabalhou com sequências de 2 a 5 palavras apenas, varrendo todo o *corpus* para a obtenção da massa de vetores de palavras. Essas sequências de n palavras são chamadas de *n-grams*.

Mesmo considerando um modelo baseado em uma sequência de poucas palavras, o treino desta rede neural recorrente continuava extremamente custoso. Mikolov (2013) propõe uma simplificação que, no lugar de obter uma saída com probabilidades das n palavras (todas as palavras do *corpus*), confrontam-se a probabilidade da palavra com apenas uma amostra de palavras, tornando o processamento bem mais rápido. Atualmente, é possível encontrar arquivos de *embeddings* pré-treinados, tanto na língua inglesa, quanto em português (ver, por exemplo, Hartmann *et al.*, 2017).

Para o experimento deste artigo, foram utilizados dois *embeddings*, um para a língua inglesa e um para a língua portuguesa. A base de vetores do Google News foi utilizada como *embedding* da língua inglesa e, assim como o *embedding* na língua portuguesa, o vetor de palavras tem dimensão 300. Para criar essa base de vetores, a fonte de textos somava em torno de 100 bilhões de palavras. Para o *embedding* na língua portuguesa, os vetores foram gerados a partir de um *corpus* tanto do português do Brasil e quanto do português europeu. Neste caso, a fonte de texto somou pouco menos de 1,4 bilhões de *tokens* (palavras válidas).

Para a obtenção de atributos para um classificador de sentimentos utilizando os vetores de palavras como parte do método é necessário relacionar a base de comentários em análise com os *embeddings* pré-treinados. A Fig. 2 ilustra dois exemplos de como obter o vetor de atributos a partir de *word embedding*.

Fig. 2 Obtenção do vetor de atributos das palavras e de vetores para utilização no classificador de sentimentos.

Cada avaliação da base de treinamento descrita em (A) deve passar por um processo de tokenização com remoção de *stopwords* como no método de *Bag of Words* e, em seguida, cada *token* é substituído pelo respectivo *embedding* (neste experimento, vetor com 300 elementos) (B). Esse é um procedimento que permite inúmeras formas de tratamento. A Fig. 2 trata de dois exemplos: a criação de um vetor de médias (C) e o chamado *bag of centroids* (D). No vetor de médias, é retirada a média de cada elemento dos vetores do *embedding* das palavras existentes em cada avaliação. Assim, cada avaliação seria formada por um novo vetor que contém as médias dos elementos do *embedding* das respectivas palavras, formando suas variáveis de entrada para o processo de classificação. No caso do *bag of centroids*, primeiro é feito um processo de agrupamento (por exemplo, *k-means*) da massa de palavras do *embedding* de acordo com seus elementos. O número de grupos é um dado configurável, mas o ideal é que não perca o poder discriminante agrupando demais as palavras e fazendo com que palavras com sentido muito diferente fiquem em um mesmo grupo. Depois deste processo, para cada comentário, verifica-se em qual grupo cada palavra pertence. O vetor de variáveis de entrada para a classificação é o número de palavras contidas na avaliação que pertencem a cada grupo, por isso o nome dado de *bag of centroids*. Por simplificação, o presente artigo apenas apresenta os resultados das análises considerando os métodos *bag of words* e o método do vetor de médias de *embeddings*.

3. BASES DE COMENTÁRIOS

Para a realização dos experimentos propostos estão sendo consideradas três bases de comentários, sendo duas delas sobre avaliações de filmes e uma sobre comentários de notícias variadas.

A base de comentários em língua inglesa contém avaliações de filmes extraídas de uma base de dados online de informação sobre música, cinema, filmes, programas e comerciais para televisão e jogos de computador, o IMDB

(utilizada por competições no Kaggle). Ela é composta por 50 mil comentários, sendo metade deles avaliações positivas e, metade, negativas. O rótulo do sentimento (positivo ou negativo) foi obtido a partir da nota dada pelo próprio usuário que escreveu a avaliação. Após aleatorização dos comentários, essa base foi dividida igualmente em base de treino e base de teste.

A base de comentários de filmes em português foi extraída de um site com um mote similar ao IMDB, no Brasil, o Filmow. Esta base é formada por aproximadamente 60 mil avaliações, sendo 40 mil utilizadas para a base de treino e 20 mil para teste. Assim como no caso da base em inglês, metade da base tem avaliações positivas e, metade, negativas. Seu rótulo foi obtido também por meio da nota dada pelo autor da avaliação.

Essas duas bases relacionadas a filmes tem características bastante distintas, não só pela língua, mas pela forma de comunicação das pessoas que estão interagindo sobre um assunto. Para a base em inglês, as avaliações eram, de maneira geral, mais densas e explicativas do que a base na língua portuguesa. Isso interfere diretamente na quantidade de informação que pode ser extraída para predizer automaticamente o sentimento de uma avaliação qualquer. Enquanto que na base do IMDB, em média, cada avaliação tem 188 *tokens*, as avaliações de filmes em português têm 10 *tokens*. De maneira geral, a base de comentários em inglês utilizada para treino de um modelo de classificação, com 25 mil avaliações, proveu 4,7 milhões de *tokens*, enquanto que a base em português, com 40 mil avaliações, proveu apenas 450 mil *tokens*.

Esse único aspecto já pode ser responsável por resultados distintos de máquinas de classificação e, por isso, o foco não é a comparação entre avaliações em português ou inglês, mas sim, como a utilização diferentes métodos de treino de modelo e a reutilização de modelos treinados e não relacionados com a base de teste pode interferir na acurácia dos resultados. Para a avaliação de acurácia de um modelo treinado em inglês, mas aplicado em avaliações em português, foi criada uma nova base de teste de aproximadamente 7 mil comentários com a tradução de comentários do português para o inglês.

Por fim, para testar o poder de predição de sentimentos em uma base de comentários cujo assunto não está correlacionado com o *corpus* utilizado no treinamento do modelo, foi levantado um número reduzido de comentários (200 comentários amostrados aleatoriamente de 15 mil comentários) de um site de notícias, o G1. Neste caso, a rotulagem do sentimento foi feita manualmente e os comentários considerados neutros foram retirados da base. Diferentemente das outras bases, o número de comentários de cunho negativo é razoavelmente maior do que comentários considerados positivos (16% apenas).

4. TÉCNICAS DE APRENDIZADO DE MÁQUINA PARA CLASSIFICAÇÃO DE TEXTOS

A partir da criação das variáveis de entrada, o problema de classificação passa a ser algo mais simples de ser resolvido.

Neste caso, cada classificador pode apresentar vantagens ou desvantagens dependendo do tipo dos dados de entrada e, portanto, não há como eleger apenas uma delas como a de melhor desempenho. Em Orimaye *et al.* (2015), por exemplo, são utilizadas as técnicas SVM (*Support Vector Machine*) e *Naïve Bayes*, enquanto que em Wawre e Deshmukh (2016) e Gupta (2016) são apresentados os resultados considerando os métodos *Random Forrest* (Floresta Aleatória), SVM e o *Naïve Bayes*. Embora existam inúmeros modelos de classificação, a classificação de sentimentos para os experimentos deste artigo foi realizada considerando os métodos: Regressão logística (RL), Floresta Aleatória (FA) e SVM linear.

5. RESULTADOS E CONCLUSÕES

Em primeiro lugar, é traçada uma linha de base desta pesquisa a partir da acurácia do modelo treinado e testado com a base em inglês nos dois métodos de construção de vetores de atributos mencionados no capítulo 2 deste artigo. Da mesma forma, serão aplicados os mesmos métodos com a base de comentários de filmes em português. Neste ponto, o objetivo é avaliar a acurácia da máquina de predição de sentimento em cada método de seleção de vetor de atributos (*bag of words* - BoW e vetor de médias com *word embeddings* - MWE), observando se as conclusões sobre a língua ou o método de classificação mostram alguma característica emergente. Os resultados deste teste são apresentados na Tabela 1.

Tabela 1. Percentual de acurácia do teste 1

Classifi- cador	Inglês		Português	
	BoW	MWE	BoW	MWE
RL	84,8	85,1	79,5	70,9
FA	83,7	83,3	78,3	69,6
SVM	82,1	85,4	78,7	71,0

Neste primeiro experimento (teste), observa-se uma clara diferença no indicador de acurácia considerando o processo de modelagem para a língua inglesa e portuguesa. Esse é um resultado esperado, já que a quantidade de informação disponível na base de treino dos comentários em inglês é maior do que em português (4,7 milhões de *tokens* de uma base contra menos de 10% na outra). Em nenhum dos casos, no entanto, o tamanho da base de treino atingiu um ponto de saturação no que tange o potencial de acurácia da predição. Uma demonstração desse fato é que, ao aumentar a quantidade de comentários de 25 mil para 43 mil da base de treino em inglês e de 40 mil para 60 mil da base de treino em português, a acurácia também aumentou. No primeiro caso, passou para 90% e, no segundo, 83%.

Outro ponto que vale notar nos resultados do primeiro experimento é que a acurácia entre métodos para a base em inglês não mostra grande discrepância. A maior diferença foi observada no classificador SVM (3,3%). No entanto, essa diferença é notável para o modelo treinado em português (menor diferença resultou em 7,7% também no classificador SVM). Uma das causas para essa diferença pode ser justamente o tamanho do *corpus* que deu origem ao *embedding* em português. Além do melhor poder de predição,

o método que utiliza o *embedding* traz maior custo computacional na fase de treinamento do modelo.

O próximo experimento tem o objetivo de avaliar a perda de acurácia ocasionada pela tradução dos comentários de filmes de português para inglês. Para isso, foram utilizados os modelos treinados no primeiro experimento para aplicação em uma amostra da base de teste. Para essa amostra, além dos comentários em português, foi criado um novo campo com a tradução destes comentários. Assim, a predição de sentimento de cada versão de um comentário (português e traduzido para o inglês) foi realizada com os respectivos modelos treinados.

Os resultados considerando o método de extração de atributos pelo BoW estão apresentados na Tabela 2 e, pelo método MWE, na Tabela 3. Os resultados de acurácia realizados completamente em inglês (treino e teste com a base IMDB) foram mantidos para facilitar a comparação.

Tabela 2. Percentual de acurácia do teste 2 - BoW

Classificador	Inglês	Português	Português traduzido
RL	84,8	80,6	70,1
FA	83,7	81,9	64,2
SVM	82,1	80,1	68,5

Neste experimento, com o método BoW, nota-se uma perda de acurácia significativa (pelo menos 10,5% de perda de acurácia) ao utilizar o modelo pré-treinado em inglês com os comentários traduzidos automaticamente do português para o inglês. Além da perda de informação com a tradução, há o fato de que apesar de ambas as bases versarem sobre avaliações de filmes, a simples mudança de ambiente virtual pode levar a diferenças nos comportamentos no uso da linguagem. Como mencionado anteriormente, a base de comentários em inglês contém uma média de palavras muito maior do que a em português e, considerando o método BoW, palavras na base de teste que não forem encontradas na base de treino não contribuem no poder de predição do modelo. No entanto, os resultados de acurácia considerando a criação de atributos pelo método MWE mostrou, pela primeira vez neste estudo, um valor melhor na acurácia em 2 dos três métodos de classificação, como mostrado na Tabela 3.

Tabela 3. Percentual de acurácia do teste 2 - MWE

Classificador	Inglês	Português	Português traduzido
RL	85,1	70,5	72,3
FA	83,3	76,2	59,6
SVM	85,4	70,5	72,1

Com o método de engenharia de atributos utilizando o *word embedding*, os resultados de acurácia tiveram um acréscimo de aproximadamente 2,2% no método de classificação RL e 3,6% SVM para a base traduzida para o inglês. Além disso, os resultados foram acrescidos em pouco mais de 1,5% (para RL e SVM) em relação ao método de *word embedding* aplicado à base não traduzida, mostrando que, apesar da perda de acurácia em utilizar um modelo treinado em uma base distinta, ainda que correlacionada, o poder do

embedding mais denso em número de *tokens* utilizados (ordem de bilhões no caso da base em inglês) trouxe ganhos no poder discriminante da máquina de classificação.

Para avaliar como os dois métodos de criação de atributos se comportam ao utilizar um modelo treinado em um contexto distinto, foi feita uma avaliação com comentários extraídos de um site de notícias em língua portuguesa. Foram extraídos em torno de 15 mil comentários e amostrados em torno de 400 deles aleatoriamente para serem submetidos à rotulagem manual (repostas possíveis: positivo, negativo e neutro). Após a rotulagem, os comentários neutros foram excluídos da base, restando 200 comentários válidos. Neste caso, a polaridade dos comentários é desbalanceada, sendo aproximadamente 84% negativos e 16%, positivos. A Tabela 4 mostra os resultados de acurácia originais da base de avaliações de filmes e dos comentários de notícias para os dois métodos de engenharia de atributos.

Tabela 4. Percentual de acurácia teste 3

Classifi- cador	Comentários de filmes		Comentários de notícias	
	BoW	MWE	BoW	MWE
RL	79,5	70,9	57,3	66,8
FA	78,3	69,6	53,3	66,3
SVM	78,7	71,0	57,8	65,3

Ao aplicar o modelo em uma base não relacionada com a que originou o modelo de classificação de sentimentos, o nível de acurácia para o método BoW ficou com uma acurácia apenas pouco maior do que uma definição aleatória da estimativa (aproximadamente 50%). No entanto, assim como observado para a base traduzida para o inglês, o método *word embedding* trouxe um ganho no indicador de acurácia em todos os classificadores, sendo que o menor ganho foi de 7,5% em relação ao percentual de acerto para o método BoW e, o maior, 13%.

Esse comportamento pode estar relacionado ao fato de que o *word embedding* carrega um significado matemático para uma quantidade de palavras, em geral, muito maior do que a base de treino do BoW. Assim, quando uma base de teste (aquela que não foi utilizada no ajuste do modelo), tiver um contexto diferente da base que deu origem ao modelo de classificação, o método que busca palavras em uma base genérica pode trazer resultados melhores na acurácia.

A acurácia por classe dos classificadores para todos os experimentos com exceção dos comentários de notícias é praticamente igual à acurácia geral. No entanto, a acurácia por classe do classificador aplicado à base de comentários em notícias pende para a classe negativa. Assim, 70% dos comentários de fato negativos foram classificados como negativos. Já dentre os comentários positivos, a acurácia tende a uma predição aleatória, com 48% de acerto.

Esse estudo apresentou uma discussão sobre dois métodos de extração de características para análise de sentimentos de avaliações em três bases de comentários. Não há, necessariamente, uma conclusão única sobre o melhor método, pois se observou ganhos nas duas abordagens ao

variar algumas situações de utilização do classificador. Novos estudos devem ser conduzidos para melhorar a acurácia dos modelos de análise de sentimentos, principalmente nos casos em que o “sentimento” a ser classificado não tem relação com os dados que geraram o modelo ajustado.

REFERÊNCIAS

- Bengio, Y., Ducharme, R., Pascal, V. e Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.
- Ding, X., Liu, B. e Yu, P.S. (2008). A holistic lexicon-based approach to opinion mining. *WSDM '08 Proceedings of the 2008 International Conference on Web Search and Data Mining*. 231-240.
- Gupta, S. (2016). Sentiment Analysis for Movie Review. *International Journal & Magazine of Engineering, Technology, Management and Research*, Vol. 3, Issue 9.
- Hartmann, N. et al.(2017). Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. *Proceedings of Symposium in Information and Human Language Technology*. Uberlandia, MG.
- Hercig,T. e Lenc, L. (2017). The Impact of Figurative Language on Sentiment Analysis. *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP*. 301–308.
- Jin, P., Zhang, Y., Chen, X. e Xia, Y. (2016). Bag-of-Embeddings for Text Classification. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*.
- Mikolov, T., Chen, K., Corrado, G. e Jeffrey Dean, J.(2013). Efficient estimation of word representations in vector space. *Proceedings of ICLR Workshop*.
- Orimaye, S.O. et al. (2015). Review-Level Sentiment Classification with Sentence-Level Polarity Correction. *CoRR abs/1511.02385*.
- Rotim, L. a Šnajder, J. (2017). Comparison of Short-Text Sentiment Analysis Methods for Croatian. *Proceedings of the 6th Workshop on Balto-Slavic Natural Language*. 69-75.
- Tang, H., Tan, S. e Cheng, X., A survey on sentiment detection of reviews. *Expert Systems with Applications*, 36, 10760–10773, 2009.
- Wawre, S. e Deshmukh, S.N. (2016). Sentimental Analysis of Movie Review using Machine Learning Algorithm with Tuned. *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 4, Issue 6.
- Yu, L.C. Wang, J., Lai, K.R. e Zhang, X. (2017). Refining Word Embeddings for Sentiment Analysis. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 534–539.