Utilizando a quantificação na análise de sentimentos em *reviews* de produtos

Daniel Zonta Ojeda[1], **Willian Zalewski**[2], **André Gustavo Maletzke**[1]

[1]Universidade Estadual do Oeste do Paraná (UNIOESTE)
Centro de Engenharias e Ciências Exatas – Foz do Iguaçu – PR – Brasil

[2]Universidade de Integração Latino-Americana (UNILA)

{daniel.ojeda, andre.maletzke}@unioeste.br, willian.zalewski@unila.edu.br

Abstract. *The constant increase in online transactions has produced an enormous amount of information, such as reviews of products. These reviews gather the consumers' sentiments about companies' products, being strategic information for companies if correctly analyzed. Understanding the quantity of positive and negative reviews about products can be explored through quantification methods. In this paper, we evaluate different quantifiers applied to product reviews and assess the influence of these methods on classification performance. We compare ten quantification methods over six product review datasets. The results indicate that eight methods outperform the naïve method for quantification tasks, and quantifiers can be employed to enhance the classification of product reviews. In both cases, statistically significant differences were observed.*

Resumo. *A coleta de informações como reviews sobre os produtos tornou-se uma tarefa relevante para as empresas, pois expressam o sentimento de consumidores sobre um determinado item. Conhecer a quantidade de reviews positivos e negativos sobre um produto/serviço é uma tarefa de interesse que pode ser explorada pela quantificação. O objetivo deste trabalho é avaliar diferentes quantificadores aplicados a reviews de produtos, bem como a influência desses métodos na performance de classificação. Foram avaliados dez métodos de quantificação em seis conjuntos de dados de reviews de produtos. Como resultado observou-se que o método amplamente utilizado para resolver tarefas de quantificação é superado por oito métodos e que quantificadores podem ser utilizados para melhorar a classificação de reviews. Em ambos os casos observou-se diferença estatisticamente significativa.*

1. Introdução

A análise de sentimentos consiste em extrair de um corpo de texto uma opinião ou sentimento relacionada a um ou mais aspectos de um produto, serviço e/ou evento expresso de forma textual [Bouazizi and Ohtsuki 2016]. O interesse crescente por essa tarefa se deve à vasta disponibilidade de dados armazenadas sobre *reviews* de produtos/serviços nos últimos anos. Para algumas tarefas de análise dessas *reviews*, o principal objetivo não é conhecer o sentimento expressado em cada *review*, mas a distribuição desses sentimentos a partir de um conjunto de *reviews*.

A solução *naïve* para este tipo de problema consiste na indução de um classificador seguido pela classificação de cada *review*. Após, para determinar a distribuição de sentimentos, é realizada a contagem do número de *reviews* classificadas como positivos e negativos. Essa estratégia é denominada de Classificar e Contar (CC) [Forman 2005]. De acordo com Forman (2005), esse método não apresenta resultados satisfatórios para uma ampla gama de aplicações, especialmente quando a distribuição de classes do conjunto de treino e de teste diferem. Embora simples, o CC tem deficiências, incluindo o erro sistêmico introduzido quando a distribuição de classes das *reviews* difere entre o conjunto de treino e teste. Para isso, Forman (2005) argumenta que problemas que envolvem predição da distribuição de classe de um conjunto de teste devem ser explorados por meio da tarefa de quantificação. A quantificação é uma tarefa proposta para estimar a distribuição de classes em uma amostra independente de dados. Ao contrário da classificação, a quantificação concentra-se na compreensão do comportamento dos grupos, em vez de reconhecer os indivíduos.

Nesse contexto, o objetivo deste trabalho é avaliar diferentes métodos de quantificação aplicados à análise de sentimentos sobre *reviews* de produtos. Adicionalmente, é apresentado como a quantificação pode aumentar a performance de classificação de sentimentos de cada *review*. Para isso, foram utilizados dez métodos de quantifcação diferentes e seis conjuntos de dados sobre *reviews* de produtos. Para cada conjunto de dados foram geradas mais de 1400 amostras com diferentes distribuições de sentimentos. As principais contribuições deste trabalho são: (*i*) apresentada envidência empírica da fragilidade do método CC na determinação da distribuição de sentimentos a partir de *reviews* de produtos; (*ii*) apresentada uma estratégia para ajustar o *threshold* de classificação com base no resultado da quantificação; e (*iii*) observada diferença estatisticamente significativa na acurácia de classificação quando o *threshold* de classificação é ajustado por métodos de quantificação.

O restante deste trabalho está organizado da seguinte maneira: na Seção 2 são apresentados conceitos básicos sobre análise de sentimentos em *reviews* de produtos e métodos quantificação; na Seção 3 são apresentados os materiais e método para o desenvolvimento deste trabalho. Nas Seções 4 e 5 são apresentados os resultados e a discussão, respectivamente; e na Seção 6 são apresentadas as conclusões e trabalhos futuros.

2. Conceitos básicos

A Análise de Sentimentos (AS) é uma abordagem computacional que se concentra na interpretação e compreensão das emoções, opiniões e atitudes expressas, em geral, mediante textos [Medhat et al. 2014]. Consiste em determinar o sentimento associado a um determinado conteúdo. Aplicações de AS envolvem problemas em diversas áreas como política mediante a análise de discurso, na educação por meio da análise de *feedbacks* de estudantes e na área de finanças com avaliações automáticas sobre o sentimento do mercado financeiro. Entretanto, nos últimos anos, têm ocorrido um aumento significativo no interesse em analisar o sentimento relacionado a opiniões de usuários sobre produtos, serviços e experiências [Tsytsarau and Palpanas 2012]. Esse interesse culminou na formalização de uma nova área denominada de Mineração de Opinião (MO). Embora, existam esforços para diferenciar as áreas de AS e MO, neste trabalho ambos os conceitos serão tratados como equivalentes.

A AS ou MO pode ser aplicada sobre níveis diferentes, os quais incluem documento, sentença e aspecto. A AS a nível de documento consiste em categorizar o sentimento geral expresso em um documento, independentemente do tamanho, isto é, pode envolver desde um capítulo de livro até um simples *review* sobre um produto. Desse modo, um documento pode ser definido de acordo com a Definição 1.

Definição 1 *(Documento) [Tsytsarau and Palpanas 2012] Um documento D é um trecho de texto expresso em linguagem natural sem restrição de tamanho.*

Um documento pode ser composto por elementos fundamentais denominados de sentenças. Na Definição 2 é apresentado o conceito de sentença.

Definição 2 *(Sentença) Uma sentença é uma unidade linguística composta por uma sequência ordenada de palavras ou tokens que encapsula uma ideia completa.*

Entretanto, em Mineração de Opinião, uma sentença ou *review* pode descrever sentimentos sobre diferentes características de um produto. Por exemplo, considere o seguinte *review*: *"O consumo de energia é ótimo, mas o preço é muito alto"*. Nesse exemplo, sentimentos com polaridades diferentes compõem a sentença, isto é, em *"O consumo de energia é ótimo"* é expresso um sentimento positivo enquanto que em *"o preço é muito alto"*, o sentimento é negativo. Portanto, o último nível de AS ou MO é baseado em aspecto. A AS em nível de aspecto consiste em categorizar os sentimentos de aspectos específicos da entidade mencionada [Bouazizi and Ohtsuki 2016]. No exemplo citado, dois aspectos são menciados sobre o produto: **consumo de energia** e **preço**, ambos com polaridades diferentes.

De acordo com [Tsytsarau and Palpanas 2012], a AS, especificamente a MO, tem ganhado maior atenção e aplicabilidade com o uso de métodos de Aprendizado de Máquina (AM), visando obter um classificador capaz de, dado um documento, sentença ou aspecto, classificar a polaridade expressa.

A classificação é uma tarefa supervisionada de AM e consiste em aprender um classificador a partir de um conjunto de treinamento $T = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, onde $\mathbf{x}_i \in \mathcal{X}$ é um vetor com m atributos no espaço de atributos $\mathcal{X}$, e $y_i \in \mathcal{Y} = \{c_1, \ldots, c_l\}$ é a classe de cada instância, respectivamente. O objetivo da classificação é prever a classe de cada exemplo na amostra de teste, isto é, um classificador é um modelo preditivo h treinado a partir de T tal que:

$$h : \mathcal{X} \longrightarrow \{c_1, \ldots, c_l\}$$

Classificadores usam vários mecanismos para atribuir a classe para um exemplo na amostra de teste. Um método é aprender um modelo que atribui pontuações (*scores*) a cada exemplo do conjunto de teste. Um *score* é um valor real que pode ser interpretado como a probabilidade posterior de uma classe, ou seja, $P(Y = c_i|\mathbf{x})$. Desse modo, um *scorer* é um modelo induzido a partir de D tal como:

$$f : \mathcal{X} \rightarrow \mathbb{R}^l$$

No contexto de AS, os dados de treinamento são documentos, representados em um espaço D, cuja dimensão é expressa por atributos extraídos de cada documento como

frequência de palavras, isto é, $D = \{(\mathbf{d}_1, s_1), \ldots, (\mathbf{d}_n, s_n)\}$, onde $\mathbf{d}_i \in \mathcal{Z}$ é um vetor com k atributos no espaço de atributos $\mathcal{Z}$, e $s_j \in \mathcal{S} = \{positivo, negativo, neutro\}$ é a classe de cada instância, respectivamente. Esses documentos podem receber como rótulo uma das seguintes classes: $positivo, negativo, neutro$. Portanto, atribuir um sentimento a um documento é definido por g, tal que:

$$g : \mathcal{D} \longrightarrow \{positivo, negativo, neutro\}, g(D_i) = \arg\max\ f(D_i, S_j)$$

Para algumas tarefas de análise de *reviews*, o principal objetivo não é conhecer o sentimento expressado em cada *review*, mas a distribuição desses sentimentos a partir de um conjunto de *reviews*. No contexto de AS, um quantificador é um modelo induzido a partir de D que prevê a prevalência de cada classe ($positivo$, $negativo$ e $neutro$) em uma amostra, tal que:

$$q : \mathcal{Q}^z \rightarrow [0, 1]^l$$

onde $\mathcal{Q}^z$ denota o universo de amostras possíveis de $\mathcal{Z}$. Dessa forma, dada uma amostra de testes $\mathcal{Q} \in \mathcal{Q}^z$, o quantificador gera um vetor $p' = [p'_1, ..., p'_l]$, onde p'_i é a probabilidade a *priori* para a classe s_i.

3. Materiais e Métodos

Para realizar a avaliação de quantificadores aplicados ao problema de análise de sentimentos em *reviews* de produtos, foram executadas as etapas apresentadas a seguir.

3.1. Etapa 1 – Preparação de conjuntos de dados

Foram construídos seis conjuntos de dados focados em produtos específicos, compostos por *reviews* de usuários em inglês. Esses conjuntos foram obtidos a partir de conjuntos de dados maiores não rotulados, dentre os quais o *Amazon Cell Phone Reviews* [Griko Nibras 2018], o *LG Refrigerator Reviews*[1] e o *Amazon Review Data* (2018) [Ni et al. 2019]. Todos os conjuntos de dados utilizados são não rotulados, ou seja, não possuem a informação sobre a polaridade do sentimento expresso no *review*. Os seguintes conjuntos de dados foram construídos:

(A) **Iphone:** possui 4950 *reviews* sobre *smartphones* da Apple, sem distinção de modelo, obtidos do *Amazon Cell Phone Reviews*;
(B) **Samsung:** possui 33610 *reviews* sobre *smartphones* da Samsung obtidos do *dataset Amazon Cell Phone Reviews*;
(C) **Xiaomi:** possui 4411 *reviews* sobre *smartphones* da Xiaomi obtidos do *dataset Amazon Cell Phone Reviews*;
(D) **Shoes:** este conjunto de dados foi obtido do *dataset* Amazon-*Shoe-Review* e contém 100.000 *reviews* de sapatos;
(E) **Games:** este conjunto de dados contém 497.577 *reviews* sobre Video Games e foi extraído do *dataset Amazon Review Data (2018)*;
(F) **LG Refrigerator:** obtido a partir do *dataset LG Refrigerator Reviews* disponível na plataforma Kaggle e contém 3.446 *reviews* de refrigeradores da marca LG.

[1] https://www.kaggle.com/datasets/vidhisrivastava/lg-refrigerator-reviews-dataset

Para converter os conjuntos de dados para um problema supervisionado foram construídas duas classes, representando sentimentos positivos e negativos a partir das notas atribuídas por cada usuário da seguinte maneira: *reviews* com nota *cinco* ou *quatro* foram considerados *reviews* que expressam sentimentos positivos e, portanto, foram rotulados como positivos. Por outro lado, os *reviews* com notas *um* ou *dois* foram considerados *reviews* que expressam sentimentos negativos e, dessa forma, foram rotulados como negativos. *Reviews* com nota *três* foram considerados neutros e descartados.

3.2. Etapa 2 – Implementação de métodos de quantificação

Nesta etapa foram implementados métodos de quantificação restritos a problemas binários. Os métodos implementados são apresentados a seguir:

***Classify and Count* (CC):** esta abordagem consiste na classificação de todas as instâncias do conjunto de teste seguida da contagem do número de exemplos pertencentes a cada classe. CC fornece resultados de quantificação ideais com um classificador perfeito. Entretanto, [Forman 2007] demonstrou que CC tem um erro sistemático que aumenta monotonicamente à medida que a distribuição do teste se afasta da distribuição de treino. Este método é dado pela Equação 1.

$$\hat{P}_{CC}(\oplus) = \frac{|\{x \in \mathcal{Q} | h(x) = \oplus\}|}{|\mathcal{Q}|} \tag{1}$$

onde $\mathcal{Q}$ denota o conjunto de teste e $\oplus$ a classe positiva;

***Adjusted Classify and Count* (ACC)** : é uma melhoria da abordagem anterior, pois permite realizar uma correção considerando taxas de verdadeiros positivos (*tpr*) e falsos positivos (*fpr*), determinadas no conjunto de treino. A abordagem ACC é dada pela Equação 2.

$$\hat{P}_{ACC}(\oplus) = \frac{\hat{P}_{CC}(\oplus) - P(\oplus|\ominus)}{P(\oplus|\oplus) - P(\oplus|\ominus)} \tag{2}$$

onde $\hat{P}_{CC}(\oplus)$ é a distribuição de classe predita pelo CC e $P(\oplus|\ominus)$ é a taxa de falsos positivos;

***Threshold Selection*:** [Forman 2007] propôs diferentes estratégias para selecionar explicitamente o *threshold* de classificação para fornecer melhores estimativas de contagem geradas pela Equação 2. As principais estratégias utilizadas são:

- **X:** seleciona o *threshold* que iguala a taxas de falsos negativos e falsos positivos (*fpr*);
- **MAX:** maximiza o denominador da Equação 2 encontrando o *threshold* que maximiza a diferença entre *tpr* e *fpr*;
- **T50:** ajusta o *threshold* para que *tpr* seja 50%;
- **Median Sweep (MS):** retorna a mediana de diversas aplicações do método ACC para um intervalo predefinido de *thresholds*.

***Distribution Matching*:** inclui métodos que misturam distribuições, geralmente distribuições de *scores* do treino com o intuito de encontrar o melhor casamento com a distribuição de *scores* do teste. O cálculo dos parâmetros dessa mistura leva à estimativa da quantificação. Foram utilizados os seguintes algoritmos:

- **HDy** [González-Castro et al. 2013]**:** este método representa as distribuições de *scores* da classe positiva e negativa mediante um

histograma. Após, uma soma ponderada desses histogramas fornece a mistura entre as distribuições de *scores* positivos e negativos, onde os pesos somam 1. Os pesos que minimizam a Distância Hellinger (HD) entre a mistura e a distribuição de *scores* do teste ($\mathcal{Q}$) são considerados a proporção das classes. O HDy é apresentado na Equação 3.

$$\hat{P}_{\mathrm{HDy}}(\oplus) = \underset{0 \leq \alpha \leq 1}{\arg\min} \{\mathrm{HD}\left(\alpha H[f(\oplus)] + (1-\alpha)H[f(\ominus)], H[f(\mathcal{Q})]\right)\} \tag{3}$$

onde $H[\cdot]$ é a operação que converte os *scores* em um histograma;

- D***y***S [Maletzke et al. 2019]**:** os autores notaram que o HDy é uma instância de um *framework* mais geral que permanecia não formalizado. Portanto, [Maletzke et al. 2019] propuseram o D*y*S, cuja formulação é apresentada na Equação 4.

$$\hat{P}_{\mathrm{DyS}}(\oplus) = \underset{0 \leq \alpha \leq 1}{\arg\min} \{\mathrm{DS}\left(\alpha R[f(\oplus)] + (1-\alpha)R[f(\ominus)], R[f(\mathcal{Q})]\right)\} \tag{4}$$

onde DS é uma medida de dissimilaridade e $R[\cdot]$ é uma operação que converte os *scores* em uma representação, como um histograma;

- **SMM** [Hassan et al. 2020]**:** é uma versão simplificada do D*y*S, porém reduzindo a complexidade de tempo. O SMM é formalizado na Equação 5.

$$\hat{P}_{SMM}(\oplus) = \underset{0 \leq \alpha \leq 1}{\arg\min} \{|\alpha\mu[f(\oplus)] + (1-\alpha)\mu[f(\ominus)] - \mu[f(\mathcal{Q})]|\} \tag{5}$$

onde $\mu[\cdot]$ representa a média de valores.

Aplicações de AS ou MO sobre *reviews* de produtos requerem conhecer o sentimento expresso não por um único *review*, mas por um conjunto de *reviews*. Essa informação possui grande relevância, pois permite medir a aceitação ou rejeição, por exemplo, de produtos, serviços e/ou campanhas. Isto é, em grande parte dos problemas de AS, a tarefa alvo não é a classificação, mas a quantificação. Porém, embora a quantificação seja a abordagem alvo, a classificação ainda permace como uma tarefa de interesse, pois analisar exemplos pontuais de sentimentos positivos e negativos pode ajudar ao tomador de decisão a compreender especificidades do problema.

Dessa forma, o resultado da quantificação pode auxiliar na performance da classificação, mediante a escolha de um melhor *threshold* de classificação. Por exemplo, dado um conjunto de *reviews* e a proporção de sentimentos positivos predita $\hat{P}(\oplus)$ por um quantificador é possível ajustar o *threshold* de classificação de modo que a diferença entre distribuição de positivos obtida classificando cada *review* e a distribuição de positivos predita pelo quantificador seja mínima. A determinação do melhor *threshold* (τ) de classificação é dada pela Equação 6.

$$\tau = \underset{0 < i < 1}{\arg\min}\, \mathbf{abs}(\hat{P}(\oplus) - |f(\mathcal{Q}) \geq i|) \tag{6}$$

3.3. Etapa 3 – Avaliação experimental

Problemas de quantificação apresentam essencialmente diferenças entre a distribuição de classes do conjunto de treinamento para o conjunto de teste. Portanto, a avaliação do desempenho de um quantificador requer que as amostras de teste reflitam a variabilidade das distribuições de classes que poderão ser observadas no mundo real [Hassan et al. 2020].

O objetivo deste trabalho é avaliar o impacto de diferentes métodos quantificação aplicado ao problema de AS em dados de *reviews* de produtos. Para isso, nesta etapa o Protocolo de Prevalência Artificial (PPA), proposto por [Forman 2005], foi aplicado para avaliar e comparar métodos de quantificação. O PPA cria vários conjuntos de testes por meio da aplicação de subamostragem. Isso significa que o PPA remove aleatoriamente exemplos da classe $\oplus$ ou $\ominus$ para gerar conjuntos de testes com distribuições de classes predeterminadas. Normalmente, são utilizadas distribuições de classes em todo o espectro de possibilidades, como $p = P(\oplus) \in \{0, .01, .02, \ldots, .99, 1\}$.

Inicialmente, os conjuntos de dados foram balanceados, utilizando a estratégia de *random undersampling*, de maneira a serem compostos do mesmo número de *reviews* positivos e negativos. Desse modo, os conjuntos de dados resultantes são compostos por 2000 instâncias com exceção do conjunto de dados LG Refrigerator que é composto por 800 instâncias. Após, cada conjunto de dados é dividido em dois subconjuntos, denominados de treino e teste, ambos com a mesma distribuição de classe. A partir do modelo pré-treinado RoBERTa e do conjunto de treino foram obtidos os *scores* de ambas as classes (positiva e negativa). Esses *scores* são utilizados pelo HDy, DyS e SMM. Também foram estimadas as taxas de verdadeiros positivos (*tpr*) e de falsos positivos (*fpr*) utilizadas nos métodos ACC, X, MAX, T50 e MS.

O conjunto de dados utilizado neste trabalho é composto por *reviews* de produtos, expressos em linguagem natural. Portanto, como classificador ou *scorer* optou-se por um modelo de linguagem pré-treinado. Para a execução dos experimentos foi utilizado o *Twitter-roBERTa-base for Sentiment Analysis*, um modelo do RoBERTa treinado em aproximadamente 58 milhões de *tweets* e ajustado para a análise de sentimentos pelo *TweetEval benchmark* [Liu et al. 2019].

Foram executados dez algoritmos de quantificação, cada um deles gerando uma estimativa de distribuição de classes. Para realizar os experimentos, foram criadas amostras do conjunto de teste, variando a distribuição da classe $\oplus$ de 0% a 100% com incrementos de 5%. Conforme observado em [Maletzke et al. 2020], o tamanho da amostra possui impacto significativo nos métodos de quantificação. Por tanto, para cada distribuição de classe foram geradas amostras de tamanho 10, 20, 30, 40, 50, 100 e 200. Para cada variação na distribuição de classe e tamanho foram geradas dez réplicas.

A avaliação dos quantificadores foi realizada usando o Erro Médio Absoluto (MAE) [Sebastiani 2020]. Para avaliar a capacidade da quantificação em influenciar a qualidade da classificação mediante a escolha do melhor *threshold* de decisão (Equação 6), foi utilizada a acurácia. A comparação estatística dos resultados foi conduzida mediante o teste de Friedman com nível de confiança de 95% e pós-teste Nemenyi. Os experimentos foram implementados na linguagem python e executados no ambiente de execução Google Colaboratory[2]. Os códigos e materiais necessários para reproduzir os experimentos realizados neste trabalho estão disponíveis no repositório Git[3].

4. Resultados

Os resultados da quantificação são apresentados na Tabela 1, na qual a primeira coluna indica os quantificadores utilizados e as demais colunas o MAE com o respectivo des-

[2]https://research.google.com/colaboratory/faq.html

[3]https://github.com/danielzontaojeda/sentiment-analysis-erbd24-paper

vio padrão para cada conjunto de dados. A última coluna é o *ranking* médio de cada quantificador. Valores em negrito indicam os melhores resultados.

Tabela 1. MAE com o respectivo desvio padrão dos quantificadores para cada conjunto de dados. Melhores resultados estão em negrito.

Dataset	Iphone	Samsung	Xiaomi	Shoes	Games	LG Refrigerator	*Ranking*
CC	0.087 (0.071)	0.066 (0.059)	0.067 (0.061)	0.079 (0.061)	0.112 (0.081)	0.128 (0.099)	8.833
ACC	0.045 (0.048)	0.039 (0.040)	0.038 (0.039)	0.055 (0.054)	0.072 (0.071)	0.052 (0.058)	3.917
MAX	0.044 (0.044)	0.038 (0.039)	0.044 (0.040)	0.055 (0.055)	0.072 (0.071)	0.063 (0.056)	4.167
T50	0.078 (0.087)	0.088 (0.090)	0.081 (0.080)	0.083 (0.087)	0.096 (0.096)	0.129 (0.118)	9.167
X	0.045 (0.046)	0.036 (0.037)	0.037 (0.037)	0.055 (0.055)	0.073 (0.071)	0.063 (0.056)	4.167
MS	**0.038 (0.042)**	**0.034 (0.036)**	**0.035 (0.036)**	**0.048 (0.049)**	**0.062 (0.064)**	0.053 (0.054)	**1.583**
HDy	0.074 (0.060)	0.062 (0.053)	0.068 (0.055)	0.064 (0.061)	0.077 (0.078)	0.171 (0.104)	8.167
SMM	0.063 (0.052)	0.050 (0.043)	0.051 (0.046)	0.069 (0.054)	0.097 (0.068)	0.105 (0.084)	7.250
SORD	**0.038 (0.041)**	**0.034 (0.035)**	**0.035 (0.035)**	0.049 (0.049)	0.063 (0.064)	**0.052 (0.053)**	1.667
DyS	0.056 (0.054)	0.048 (0.044)	0.051 (0.045)	0.056 (0.056)	0.071 (0.074)	0.137 (0.091)	6.083

O quantificador MS apresentou melhor resultado na predição de $P(\oplus)$ em cinco dos seis conjuntos de dados. O SORD empatou com o MS em três conjuntos de dados, superando o MS no conjunto de dados LG Refrigerator, assumindo o segundo melhor *ranking* entre os quantificadores avaliados.

Após, foi avaliado impacto da quantificação na classificação. Para isso, após a quantificação de cada amostra de teste, o *threshold* de classificação do RoBERTa foi ajustado conforme Equação 6. Os resultados da classificação de cada amostra com o auxílio dos quantificadores são apresentados na Tabela 2. A primeira coluna indica o quantificador utilizado e as demais colunas, com exceção da última, os valores médios das acurácias com o respectivo desvio padrão. A última coluna representa o *ranking* médio do classficador RoBERTa quando o resultado da quantificação, para cada quantificador, é utilizado para ajustar dinamicamente o *threshold* de decisão.

Tabela 2. Acurácia média e desvio padrão da classificação ajustada baseada no resultado da quantificação para cada conjunto de dados. Melhores resultados estão em negrito.

Dataset	Iphone	Samsung	Xiaomi	Shoes	Games	LG Refrigerator	Rank médio
CC	0.893 (0.070)	0.921 (0.059)	0.924 (0.060)	0.869 (0.064)	0.819 (0.075)	0.853 (0.095)	9.000
ACC	0.917 (0.060)	0.936 (0.048)	0.942 (0.048)	0.892 (0.070)	0.855 (0.081)	0.885 (0.073)	5.250
MAX	0.920 (0.056)	0.938 (0.047)	0.938 (0.049)	0.893 (0.067)	0.855 (0.080)	0.891 (0.069)	4.333
T50	0.894 (0.083)	0.896 (0.087)	0.905 (0.076)	0.870 (0.090)	0.839 (0.097)	0.844 (0.112)	9.167
X	0.918 (0.057)	0.940 (0.047)	0.942 (0.048)	0.893 (0.070)	0.857 (0.082)	0.891 (0.069)	3.417
MS	0.922 (0.058)	**0.939 (0.046)**	0.944 (0.047)	**0.912 (0.067)**	**0.862 (0.080)**	0.892 (0.070)	1.833
HDy	0.908 (0.067)	0.923 (0.056)	0.922 (0.059)	0.889 (0.073)	0.857 (0.085)	0.822 (0.106)	7.583
SMM	0.905 (0.058)	0.928 (0.049)	0.931 (0.054)	0.876 (0.063)	0.829 (0.070)	0.871 (0.083)	7.500
SORD	**0.923 (0.058)**	**0.939 (0.046)**	**0.945 (0.047)**	0.896 (0.067)	**0.862 (0.080)**	**0.893 (0.069)**	**1.500**
DyS	0.917 (0.062)	0.932 (0.052)	0.934 (0.052)	0.893 (0.070)	0.859 (0.083)	0.852 (0.093)	5.417

Após, os resultados das acurácias de classificação foram comparados utilizando o Teste de Friedman com pós-teste Nemenyi. Na Figura 1 é apresentado o diagrama de diferença crítica. Os quantificadores conectados pela barra horizontal são os que não apresentam diferença estatisticamente significativa entre si.

Dos dez métodos avaliados, nove apresentaram resultados melhores em relação ao CC, sendo que três (MS, SORD e X) com diferença estatisticamente significativa.

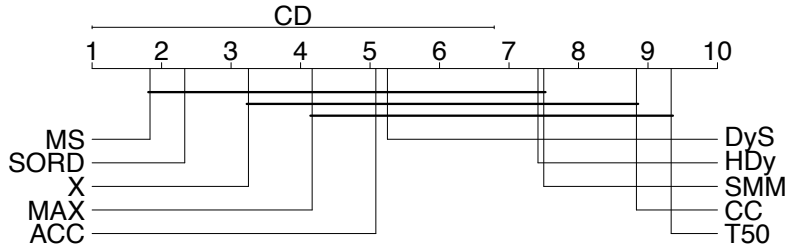

Figura 1. Diagrama de diferença crítica das acurácias de classificação.

5. Discussão

A análise de sentimentos expressadas por consumidores mediante *reviews* constitui informação essencial para empresas que desejam entender seus consumidores e a interação de seus produtos e/ou serviços com os mesmos. Em geral, analisar esses *reviews* de forma individual possui pouca relevância, pois o que se busca é entender os principais pontos positivos e negativos do produto e/ou serviço com base em diversos *reviews*, isto é, determinar a distribuição da polaridade de sentimentos sobre aquele produto. Para isso, a tarefa adequada não é a classificação, mas a quantificação.

Entretanto, a quantificação ainda é incipiente na área de análise de sentimento voltada a *reviews* de produtos, sendo que, em geral, as aplicações que exploram *reviews* de clientes, utilizam o método naïve de Classificar e Contar (CC). Os resultados da Tabela 1, evidenciam a fragilidade do método CC, o qual foi superado por oito dos nove métodos de quantificação avaliados. Portanto, aplicações que buscam responder a perguntas como: "*quantos clientes gostaram da nova funcionalidade do produto*" ou "*após a nova campanha de marketing qual foi a positividade das avaliações*", serão melhor respondidas com a utilização de métodos de quantificação.

Outro ponto positivo da quantificação em aplicações que envolvam a classificação de *reviews* é evidenciado na Tabela 2. Nela, são apresentadas as acurácias de classificação de cada *review* que compõem as amostras extraídas do conjunto de teste, utilizando o modelo de linguagem RoBERTa como classificador, porém ajustando o *threshold* de decisão conforme o resultado da quantificação de cada quantificador. Os resultados de acurácia sem ajuste do *threshold*, isto é, usando o método CC, são inferiores a todos os resultados em que algum método de quantificação foi utilizado com exceção do T50. Além disso, três dos métodos de quantificação quando utilizados para melhorar a classificação apresentaram resultados estatísticamente significativos, sendo os métodos MS e SORD os que apresentaram melhores resultados.

A quantificação, embora ainda pouco conhecida pela comunidade de aprendizado de máquina e mineração de dados, apresenta-se como um ferramenta promissora tanto para atacar problemas que são, em essência, de quantificação como problemas que envolvam a classificação de *reviews*.

6. Conclusões

Embora a estratégia de classificar e contar seja o método comumente utilizado, neste trabalho mostrou-se empiricamente que essa estratégia é um *baseline*, sendo superada estatisticamente por outros métodos na tarefa de quantificação de sentimentos em *reviews*

de produtos. Demonstrou-se empiricamente que a quantificação pode desempenhar um papel auxiliar na classificação, determinando um novo *threshold* de decisão entre *reviews* positivas e negativas. Trabalhos futuros incluem a avaliação de outros quantificadores e conjuntos de dados, bem como a extração e quantificação de aspectos dos *reviews*.

Referências

Bouazizi, M. and Ohtsuki, T. (2016). Sentiment Analysis in Twitter: From Classification to Quantification of Sentiments within Tweets. In *2016 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, Washington, DC, USA. IEEE.

Forman, G. (2005). Counting Positives Accurately Despite Inaccurate Classification. In *Machine Learning: ECML 2005*, volume 3720, pages 564–575. Springer Berlin Heidelberg, Berlin, Heidelberg.

Forman, G. (2007). Quantifying counts, costs, and trends accurately via machine learning. Technical report, Technical report, HP Laboratories, Palo Alto, CA.

González-Castro, V., Alaiz-Rodríguez, R., and Alegre, E. (2013). Class distribution estimation based on the Hellinger distance. *Information Sciences*, 218:146–164.

Griko Nibras (2018). Amazon Cell Phones Reviews. https://www.kaggle.com/datasets/grikomsn/amazon-cell-phones-reviews.

Hassan, W., Maletzke, A., and Batista, G. (2020). Accurately Quantifying a Billion Instances per Second. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10, Australia. IEEE.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Maletzke, A., Dos Reis, D., Cherman, E., and Batista, G. (2019). DyS: A Framework for Mixture Models in Quantification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):4552–4560.

Maletzke, A., Hassan, W., dos Reis, D., and Batista, G. (2020). The Importance of the Test Set Size in Quantification Assessment. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 2640–2646, Japan.

Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Ni, J., Li, J., and McAuley, J. (2019). Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, China.

Sebastiani, F. (2020). Evaluation measures for quantification: An axiomatic approach. *Information Retrieval Journal*, 23(3):255–288.

Tsytsarau, M. and Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24:478–514.