Atividades Unidade 3 - Processamento de Linguagem Natural

Aluno: Carlos Eduardo Falandes

01/12/2024

TC.3.2. Qual a relação entre as etapas de pré-processamento de texto e a redução de dimensionalidade quando se lida com extração de características? Ilustre isso considerando 2 exemplos que façam uso de stemização e/ou lematização.

A principal relação entre as etapas de pré-processamento do texto e a redução de dimensionalidade é que o pré-processamento simplifica os dados textuais, ele remove elementos que podem dificultar a interpretação do texto ou até mesmo torná-la impossível. Ao realizar essa simplificação, a etapa elimina redundâncias e variações desnecessárias, facilitando a extração de características e tornando o conjunto de dados menos complexo para a etapa de aprendizado de máquina.

A stemização e a lematização são fundamentais para reduzir a dimensionalidade do conjunto de dados, pois ambas diminuem significativamente a quantidade de palavras únicas no vocabulário.

- **Stemização:** Reduz as palavras ao seu radical, ignorando flexões ou terminações.
- **Lematização:** Simplifica as palavras à sua forma base, removendo variações gramaticais enquanto preserva o significado.

Exemplo 1: Original: *"Os produtos são incríveis! Gostei muito da qualidade oferecida."*
Com stemização: *"prod incriv gost muit qualidad oferec"*

Exemplo 2: Original: *"A aprendizagem profunda tem aplicações robustas em visão computacional."*
Com lematização: *"aprendagem profundo aplicação robusto visão computacional"*

PP.4.2. Construa um modelo do tipo word2vec (W2V) para classificação de revisões de produto que atenda aos seguintes critérios:

a) O modelo deve operar sobre dados de revisão que tenham sido pré-processados;
b) O modelo deve ser comparado, em termos de desempenho de classificação, com um modelo clássico do tipo bag of words (BOW) com transformação TFIDF.
c) Para a classificação, utilizar no mínimo 15 reviews de treinamento, classificador utilizando Multilayer Perceptron e mais 45 reviews de validação, sendo elas igualmente distribuídas entre revisões positivas, negativas e neutras.

Código da atividade: `Product Review Classifier with IA`

Table 1: Métricas calculadas para Word2Vec e BOW-TFIDF

Métrica	**Word2Vec**	**BOW-TFIDF**
T (*Total samples*)	45.00	45.00
TPR (*True Positive Rate*)	0.60	0.80
FPNR (*False Positive Negative Rate*)	0.33	0.07
FNeP (*False Neutral Positive Rate*)	0.00	0.01
FPR (*False Positive Rate*)	0.10	0.24
TNR (*True Negative Rate*)	0.60	0.47
FNP (*False Negative Positive Rate*)	0.20	0.47
FNNe (*False Negative Neutral Rate*)	0.20	0.07
FNR (*False Negative Rate*)	0.17	0.10
TNeR (*True Neutral Rate*)	0.73	0.80
FNeN (*False Neutral Negative Rate*)	0.00	0.13
FNeR (*False Neutral Rate*)	0.04	0.07
Acurácia (*Accuracy*)	0.64	0.69
Verossimilhança positiva	5.87	3.36
Verossimilhança negativa	3.60	4.67
Verossimilhança neutra	18.33	11.61

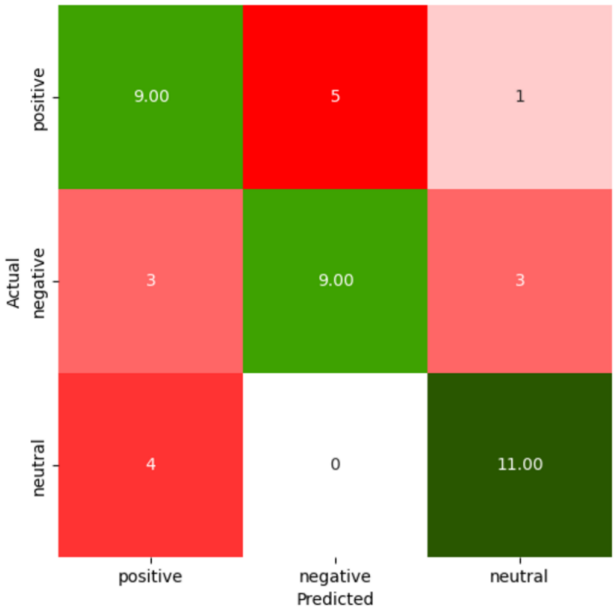

Figure 1: Matriz de confusão Word2Vec

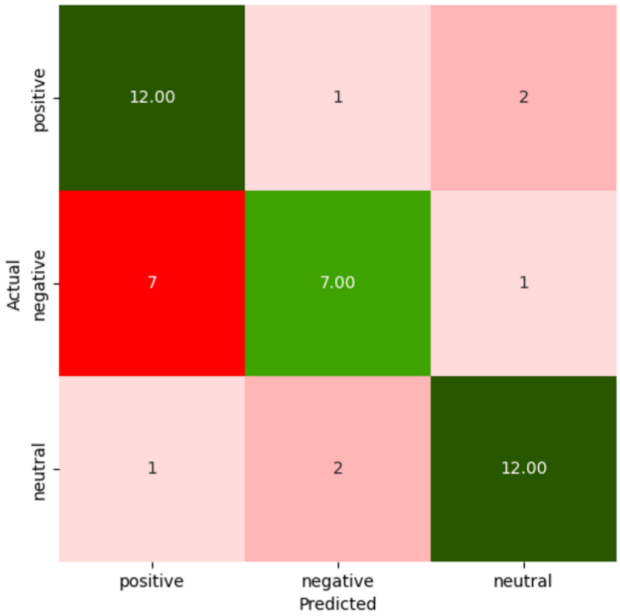

Figure 2: Matriz de confusão BOW-TFIDF

Analisando os dados obtidos pelo seu modelo e os valores para as várias taxas, acurácia geral e razões de verossimilhança, diga o que poderá acontecer caso esse modelo seja utilizado em uma base de dados contendo novas reviews (qual a previsão de comportamento da classificação)?

As métricas fornecidas para ambos os modelos Word2Vec e BOW-TFIDF, torna possível fazer algumas previsões sobre o comportamento da classificação se esses modelos forem utilizados em uma base de dados contendo novas reviews:

Word2Vec:

1. **Acurácia:** 0.64 (64%) – A acurácia geral indica que o modelo acerta 64% das classificações. Essa taxa não é tão alta, sugerindo que o modelo pode estar cometendo erros em 36% dos casos.

2. **Taxa de Verdadeiros Positivos (TPR):** 0.60 (60%) – O modelo está conseguindo identificar corretamente 60% das instâncias positivas (ou seja, aquelas que pertencem à classe positiva).

3. **Taxa de Falsos Positivos (FPR):** 0.10 (10%) – O modelo tem uma taxa relativamente baixa de falsos positivos, o que significa que ele raramente classifica incorretamente uma instância negativa como positiva.

4. **Taxa de Falsos Negativos (FNR):** 0.17 (17%) – Embora seja moderada, uma taxa de falsos negativos de 17% sugere que o modelo pode deixar passar uma quantidade considerável de amostras positivas.

O modelo Word2Vec provavelmente terá uma boa performance em identificar amostras positivas, mas poderá errar ao classificar amostras negativas e neutras. Isso pode indicar que, em uma base de dados com novas *reviews*, o modelo pode ser mais eficaz em classificar as *reviews* positivas corretamente, mas terá dificuldades em lidar com o equilíbrio de classes, especialmente com a classe negativa.

BOW-TFIDF

1. **Acurácia:** 0.69 (69%) – O modelo tem uma acurácia ligeiramente superior ao Word2Vec, o que indica uma melhor performance geral.

2. **Taxa de Verdadeiros Positivos (TPR):** 0.80 (80%) – O modelo é muito mais eficaz em identificar instâncias positivas, com 80% de acerto.

3. **Taxa de Falsos Positivos (FPR):** 0.24 (24%) – A taxa de falsos positivos é mais alta em comparação com o modelo Word2Vec, o que significa que o modelo pode ter mais dificuldades em não classificar amostras negativas como positivas.

4. **Taxa de Falsos Negativos (FNR):** 0.10 (10%) – A taxa de falsos negativos é mais baixa que a do modelo Word2Vec, o que sugere que ele consegue identificar mais amostras positivas corretamente.

O modelo BOW-TFIDF terá uma performance muito boa ao identificar amostras positivas (80% de TPR), mas com uma taxa de falsos positivos mais alta (24%). Isso significa que, em uma base de novas reviews, o modelo pode acabar classificando algumas amostras negativas como positivas, o que pode levar a uma classificação incorreta, especialmente se houver uma alta proporção de reviews negativas. Contudo, o modelo tende a ser mais eficaz na identificação de amostras positivas e menos propenso a cometer falsos negativos do que o modelo Word2Vec.

O que acontece se o número de amostras de uma das classes for reduzido enquanto o número de amostra das demais for mantido? (ex. Utilize 2 amostras negativas, 8 positivas e 8 neutras para treinamento e repita o cálculo das taxas de acerto e erro para positivas, negativas e neutras, assim como os demais indicadores de desempenho – acurácia e razões de verossimilhança).

Table 2: Métricas calculadas para Word2Vec com dados de treinamento não equiparados e equiparados

Métrica	**Word2Vec**	**BOW-TFIDF**
T (*Total samples*)	45.00	45.00
TPR (*True Positive Rate*)	0.40	0.67
FPNR (*False Positive Negative Rate*)	0.00	0.00
FNeP (*False Neutral Positive Rate*)	0.04	0.02
FPR (*False Positive Rate*)	0.15	0.31
TNR (*True Negative Rate*)	0.07	0.13
FNP (*False Negative Positive Rate*)	0.27	0.60
FNNe (*False Negative Neutral Rate*)	0.67	0.27
FNR (*False Negative Rate*)	0.03	0.00
TNeR (*True Neutral Rate*)	0.73	0.93
FNeN (*False Neutral Negative Rate*)	0.07	0.00
FNeR (*False Neutral Rate*)	0.32	0.18
Acurácia (*Accuracy*)	0.40	0.58
Verossimilhança positiva	2.61	2.14
Verossimilhança negativa	2.00	∞
Verossimilhança neutra	2.28	5.32

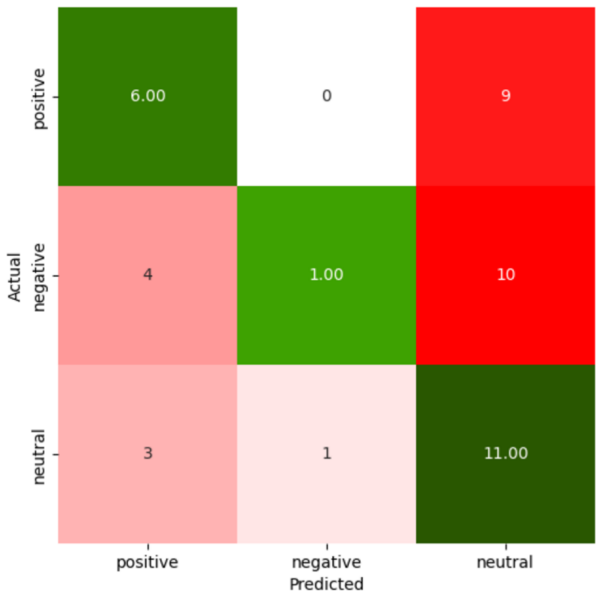

Figure 3: Matriz de confusão Word2Vec para amostras de treinamento não equiparadas

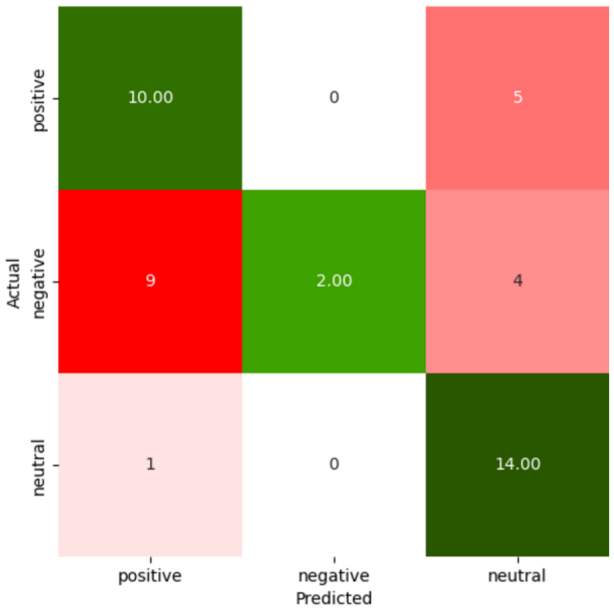

Figure 4: Matriz de confusão BOW-TFIDF para amostras de treinamento não equiparadas

Quando o número de amostras de uma das classes é reduzido enquanto o número de amostras das demais é mantido, isso geralmente afeta o desempenho do modelo, especialmente nas classes com menos amostras. Sendo assim pode-se observar como isso impacta cada métodos.

Word2Vec

1. **Acurácia:** 0.40 (40%) – A acurácia caiu significativamente, indicando que o modelo tem dificuldades gerais em classificar corretamente com a distribuição desbalanceada.

2. **Taxa de Verdadeiros Positivos (TPR):** 0.40 (40%) – A TPR caiu, sugerindo que o modelo teve dificuldade em identificar corretamente as instâncias positivas (em comparação com 60% antes).

3. **Taxa de Verdadeiros Negativos (TNR):** 0.07 (7%) – A TNR também caiu drasticamente, indicando que o modelo tem dificuldade em identificar corretamente as instâncias negativas, provavelmente devido à escassez de amostras negativas.

4. **Taxa de Falsos Positivos (FPR):** 0.15 (15%) – A FPR aumentou, indicando que o modelo classificou mais amostras negativas como positivas.

5. **Taxa de Falsos Negativos (FNR):** 0.03 (3%) – A FNR diminuiu, indicando que o modelo cometeu menos erros ao deixar de identificar amostras positivas.

6. **Taxas de Likelihood Ratio:** A razão de verossimilhança para amostras negativas caiu para 2.00, o que indica que a classificação de amostras negativas é menos confiável devido à baixa quantidade de amostras dessa classe.

A redução do número de amostras negativas fez com que o modelo tivesse um desempenho significativamente pior na identificação de amostras negativas e, de forma geral, resultou em uma acurácia consideravelmente mais baixa. A escassez de amostras negativas também causou um aumento no número de falsos positivos e uma taxa de verdadeiros negativos muito baixa.

BOW-TFIDF

1. **Acurácia:** 0.58 (58%) – A acurácia caiu em relação ao valor anterior de 0.69, mas ainda é superior ao modelo Word2Vec. Isso demonstra que o BOW-TFIDF mantém um desempenho relativamente bom, embora a redução de amostras negativas tenha impactado sua performance.

2. **Taxa de Verdadeiros Positivos (TPR):** 0.67 (67%) – A TPR diminuiu em relação aos 80% anteriores, mas o modelo ainda consegue identificar bem as amostras positivas.

3. **Taxa de Verdadeiros Negativos (TNR):** 0.13 (13%) – A TNR sofreu uma redução significativa, indicando dificuldades na identificação correta de amostras negativas após a diminuição de amostras dessa classe.

4. **Taxa de Falsos Positivos (FPR):** 0.31 (31%) – A FPR aumentou, mostrando que o modelo está cometendo mais erros ao classificar amostras negativas como positivas.

5. **Taxa de Falsos Negativos (FNR):** 0.00 (0%) – A FNR permanece zerada, indicando que o modelo continua identificando todas as amostras positivas corretamente.

6. **Taxas de Likelihood Ratio:** ∞ – A razão de verossimilhança para amostras negativas tornou-se "infinita", refletindo a extrema dificuldade do modelo em classificar corretamente amostras negativas devido à escassez de exemplos dessa classe.

A escassez de amostras negativas afetou gravemente a capacidade do modelo BOW-TFIDF de classificar corretamente instâncias negativas. Embora o modelo continue a identificar bem as instâncias positivas, ele comete muitos falsos positivos e tem extrema dificuldade com as instâncias negativas, o que é refletido na taxa "infinita" de likelihood ratio para as negativas.

A redução do número de amostras de uma classe (negativa, neste caso) provoca um desbalanceamento nas classes, afetando negativamente a capacidade do modelo de classificar corretamente as instâncias dessa classe. Ambas as métricas (Word2Vec e BOW-TFIDF) mostram uma queda na acurácia, taxa de verdadeiros negativos (TNR) e um aumento na taxa de falsos positivos (FPR). O modelo BOW-TFIDF parece sofrer ainda mais com a escassez de amostras negativas, enquanto o Word2Vec apresenta uma queda mais generalizada nas taxas de desempenho.