A base de dados Reuters 21578 reúne notícias organizadas em 135 categorias, havendo alguns documentos, embora poucos, que aparecem em mais de uma classe.

Muitos pesquisadores optam por trabalhar com apenas um subconjunto das classes disponíveis na coleção, o que compromete a comparação dos resultados. Uma maneira de tratar esse problema é utilizar partições no modo APTÉ (ModApté split).

Existem 3 maneiras de se fazer isso:

1. Utilizar o conjunto formado pelas 10 categorias com a maior quantidade de amostras positivas para treinamento. Trata-se do R10
2. Utilizar o conjunto formado pelas 90 categorias com pelo menos uma amostra positiva para treinamento e pelo menos uma amostra positiva para teste. Trata-se do R90
3. Utilizar o conjunto formado pelas 115 categorias com pelo menos uma amostra de treinamento. Trata-se do R115

No modo Apté, o que tem sido mais frequente é a seleção de 9603 documentos para treinamento e 3299 documentos para teste.

Referências:

[1] Nardiello P., Sebastiani F., Sperduti A. (2003) Discretizing Continuous Attributes in AdaBoost for Text Categorization. In: Sebastiani F. (eds) Advances in Information Retrieval. ECIR 2003. Lecture Notes in Computer Science, vol 2633. Springer, Berlin, Heidelberg

[2] Pietramala A., Policicchio V.L., Rullo P., Sidhu I. (2008) A Genetic Algorithm for Text Classification Rule Induction. In: Daelemans W., Goethals B., Morik K. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2008. Lecture Notes in Computer Science, vol 5212. Springer, Berlin, Heidelberg

[3] Mountassir A., Benbrahim H., Berrada I. (2014) The Nearest Centroid Based on Vector Norms: A New Classification Algorithm for a New Document Representation Model. In: Perner P. (eds) Machine Learning and Data Mining in Pattern Recognition. MLDM 2014. Lecture Notes in Computer Science, vol 8556. Springer, Cham

IMPORTANTE: Todas as referências aqui citadas utilizaram a mesma base do Reuters que nós: Reuters-21578 text categorization test collection (Distribution 1.0, David D. Lewis)

[1] Tudo o que sei é que utilizaram o R115. (Não consegui acesso ao arquivo full).

[2] “A Reuters-21578 é composta por 12.902 documentos. Usamos a divisão ModApte, na qual 9.603 documentos são selecionados para treinamento (dados vistos) e os outros 3.299 formam o conjunto de testes (dados não vistos). Das 135 categorias do grupo TOPICS, consideramos o 10 com o maior número de exemplos positivos de treinamento (a seguir, vamos nos referir a este subconjunto como R10). Enfatizamos que usamos todos os 9603 documentos do corpus de treinamento para a fase de aprendizado e realizamos o teste usando todos os 3299 documentos do conjunto de teste (incluindo aqueles que não pertencem a nenhuma categoria no R10).”

[3] Utilizaram a base do Reuters com os 9603 documentos para treinamento e 3299 para testes.

https://miguelmalvarez.com/2015/03/20/classifying-reuters-21578-collection-with-python-representing-the-data/