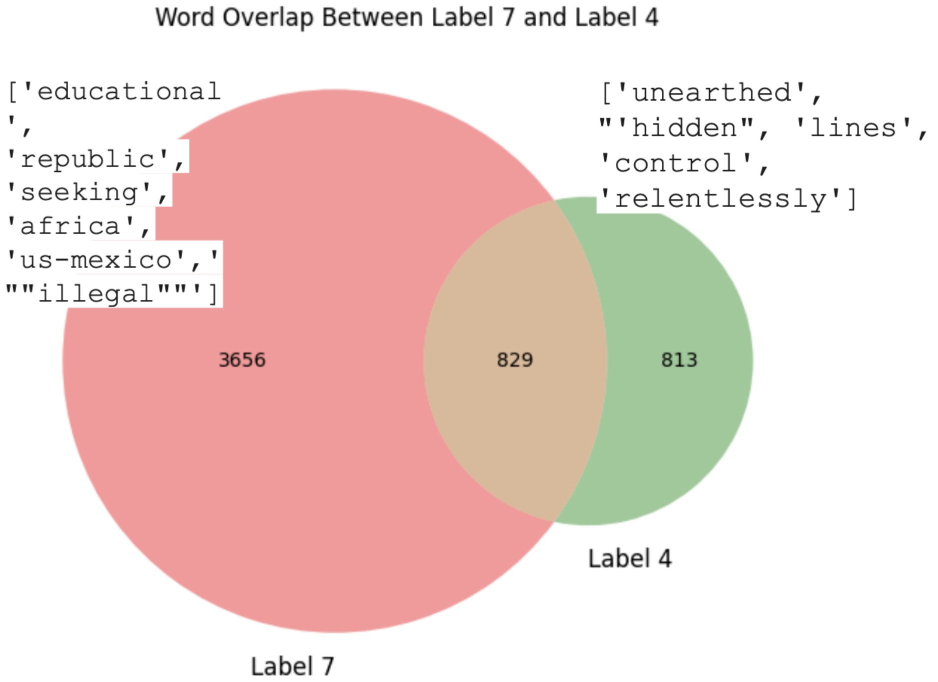
Analysis:

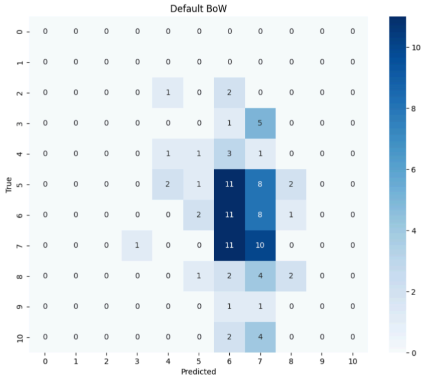
Word overlap between articles with reliability scores of 4 and 7 is depicted in the Venn diagram. Even though the two categories share 829 words, a significant portion of the words, 3656 for label 7 and 813 for label 4 remain unique to each category, indicating significant linguistic differences in the framing of these articles. Terms like "educational," "republic," "us-mexico," and "illegal" are frequently used in Label 7 articles, suggesting an emphasis on institutional or political themes that may be connected to government narratives or legal discourse. Words like "unearthed," "hidden," "control," and "relentlessly," on the other hand, are used in label 4 articles and convey a tone that is more emotionally charged or investigative.

This imbalance directly affects a majority class classifier, which always predicts the most common label. Regardless of the actual tone or content of an article, the classifier would probably default to label 7 because it dominates both in frequency and word diversity. Articles that use language consistent with label 4 (emotive or investigative), may therefore be routinely misclassified. When the objective is to distinguish small framing differences related to perceived reliability, this illustrates how majority class classification can obscure significant linguistic distinctions and perform poorly.

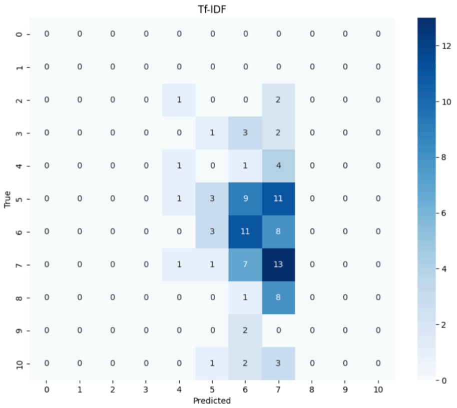


The majority of predictions are grouped around reliability scores of 5-7, according to the confusion matrices. Based on our annotation guidelines these scores indicate moderately reliable articles with partial balance, emotional language that is present but not overpowering, and a slightly suggestive tone. These articles are more difficult to categorize accurately because they often contain elements of neutrality along with emotional language or bias. Because assessing tone, balance, and emotive language in immigration news is subjective and ambiguous, the models frequently misclassify the perceived reliability by one point (predicting 5 or 7 instead of 6). The TF-IDF model exhibits somewhat tighter clustering and fewer extreme misclassifications than Bag of Words, indicating a slight improvement in capturing the subtleties of perceived media reliability.

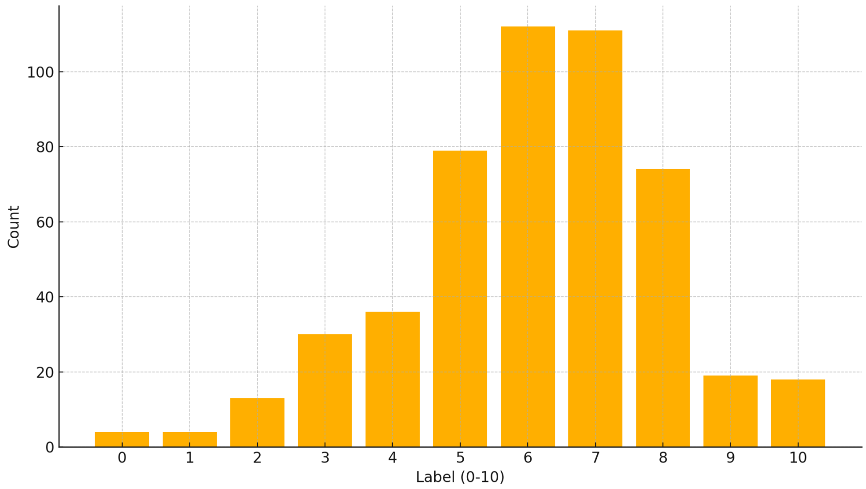
Bag of Words:
0.240, 95% CIs: [0.156 0.324]



TF-IDF:
0.280, 95% CIs: [0.192 0.368]



Based on our final adjudicated annotations, most of the labels fall between reliability scores of 5-8, which closely aligns with the patterns observed in the confusion matrices. This seems to suggest that our model has likely learned this distribution, as it tends to mainly predict scores in the middle range.


Count
Label (0-10)