False Negatives Analysis in Recall Estimation of NLP Models – White Paper

Keqiao Li

1. **Project Summary**

This white paper summarized all the findings and takeaways from the intern project, including:

1. False Negatives (FNs) analysis in Recall Estimation (RE) for all 4 CIB Synthesys NLP models: Complaints, Cheating & Illegality (Candi), Rumors and Secrecy.
2. FNs analysis for NLP models trained on open-source Consumer Complaint Dataset from Consumer Financial Protection Bureau (CFPB).
3. TBD.

The overall conclusion of the project is that the assumption of stratified sampling in RE for NLP models is potentially violated and not performed as expected, which could cause larger uncertainties in RE for the model monitoring purpose.

1. **Motivation**

Monitoring and maintaining the deployed Machine Learning (ML) models play significant roles in the Lifecyle of all ML models. The most commonly used metrics are precision and recall. It is easy to estimate precision by randomly sampling from the population of sentences and query for labels from human annotators to mark incorrectly classified instances. However, recall estimation is not as straightforward to use the same approach. To estimate recall, one has to look at the entire universe of the sentences, which is not feasible in production given the large number of sentences (over 10M) and rare alerts.

Random sampling is a straightforward sampling method without any assumptions. However, it requires samples to be labeled which is beyond SME labeling capability. For instance, to estimate with a Margin of Error (MOE) of 0.01% at 95% confidence level, 38412 samples are needed to be labeled, which is not feasible in production. Currently, stratified sampling approach is applied to estimate recall in production under the assumptions of:

1. Most of FNs are close to the threshold (concentrated in the high-risk bin, e.g., 800 out of 1000).
2. There are a few FNs in the moderate-risk bin (e.g., 190 out of 1000).
3. There are negligible number of FNs in the low-risk bin (e.g., 10 out of 1000, or aggressively say, zero FNs).

Therefore, the low-risk bin can be considered as ***unsampleable*** because of negligible number of FNs but vast majorities of sentences, which results in a small (e.g., or even smaller). On the contrary, moderate-risk bin and high-risk bin can be considered as ***sampleable*** because of vast majorities of FNs but a small number of sentences, which results in a relatively large (e.g., or even larger). Sampling from the ***unsampleable*** bin will run into the same problem as random sampling does. Hence, by only sampling from the ***sampleable*** bins and ignoring the ***unsampleable*** bins, FNs can be sampled with higher chance so that the estimated recall will be more accurate with less uncertainties.

However, the analysis shows the whole range, , is ***unsampleable***. The FNs are distributed across the range of . Even worse, FNs distribution could be highly right skewed, which means the vast majorities of FNs are concentrated in the range close to zero. As a result, is small everywhere in , e.g., or even smaller.

1. **FNs analysis**

FNs by 4 NLP models (Complaints, Candi, Rumors and Secrecy) on 1-week data from Q1 2022 has been analyzed. Figure xx shows a very typical example of the range of is ***unsampleable*** and the assumption of stratified sampling is potentially violated. The distribution of FNs is highly right skewed, where the vast majorities of FNs are concentrated in , and the rest of FNs are spread out in . Such right skewed distribution provides evidence of the contradiction with the assumption of stratified sampling approach. Overall, there are ~55% of FNs in and ~24% of FNs at the value of 0.5. Only ~13% and ~10% of FNs are in and , respectively (see Table 1).

Table 2 shows recall estimation using stratified sampling approach is statistically not feasible in production. For instance, bin 0-0.1, over 4 million samples are needed given the MOE of 0.0001% at 95% confidence level. A smaller MOE may even be required, which means more samples are needed.

FNs rate is plotted as the function of probability score in Figure xx, which is weakly monotonically increasing. This observation validates the speculation about the monotonicity of FNs rate in stratified sampling for NLP models as other teams interpolated the FNs rate using linear regression. It is noticed that subjectively defining bins can give us strict monotonicity.

1. **CFPB dataset**

The above FNs analysis is conducted without knowing the ground truth label of the entire universe of sentences. However, it is important to have a good understanding of the “true” FNs distribution. Therefore, we use open-source Consumer Complaint Dataset from CFPB hoping that replicating the similar FNs distribution can be helpful in supporting our conclusion from the above FNs analysis. The model development process includes the following steps: data preparation, preprocessing, model pipeline, hyperparameter tuning and finding the threshold.

1. Data preparation: consumer Complaint Dataset (01/01/2018-07/01/2022) is downloaded from the CFPB, and manually assigned all data categorized into “money transfer” to be label “1”, and others to be label “0”. It ends up with 2,831 positives and 162,700 neutral in the training dataset, and 4,011 positives and 439,891 neutral in the test dataset.
2. Preprocessing: stop words (e.g., “XX/XX/2020”) have been removed and all tokens are lemmatized.
3. Model pipeline: CountVector + SelectKBest + LogisticRegression with intercept set to be *False*.
4. Hyperparater tuning: using Gridsearchcv to select best hyperparameter combination based on the “average precision” metric because of its rubustness to threshold movement.
5. Finding the threshold: using precision-recall curve to find the threshold.

The FNs in Figure XX has a similar distribution to the one in Figure 1, which is contracdicted to the assumption of stratified sampling. Over 57% of FNs are in , and there are ~32% of FNs in and ~11% of FNs in , where 0.76 is the threshold. Under the MOE of 0.01% at 95% confidence level, it requires 294,982 samples to estimate FN rate for bin 0-0.1, which is, again, beyond the SME labeling capability. Figure XX shows monotonic increasing of FNs rate as a function of proabbiltiy score.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

1. **Motivation**

Model monitoring requires a sound understanding of the key metrics of the model performance. While precision can be easily and inexpensively measured by sampling from the classified sentences and unitizing (paid) human computation to mark incorrectly classified instances, it is not as straightforward to use the same approach for measuring recall. In the case of recall, one needs to sample from the entire universe of the sentences data. However, positives are sparse which leads the dataset is extremely imbalanced. Given the limited labeling capability of annotators, at certain confidence level with a small margin of error, orders of magnitude more data need to be sampled if the positives are rare. Consequently, different sampling strategies has been explored. For example, random sampling is preferred because of its simplicity and assumption free. However, this is not feasible in production due to the rare number of positives. On the other hand, stratified sampling is considered but with some assumptions (blue inverse triangle in Figure 2): (1) Vast majorities of FNs are close to the threshold(e.g., 0.95 in Figure 2). Such area is so-called high risk bin (2) Only A few FNs in the low risk bin(e.g., the range from 0.5-0.85 in Figure 2). (3) Almost zero FNs in the area close to 0. Such assumption allows us to only focus on randomly sampling potential FNs from a relative small chunk of dataset. Within such area, the FNs proportion to the local population is high so that FNs can be sampled with a high probability.

Figure 2. Stratified sampling assumption illustration

1. **False Negatives Analysis**

However, our analysis found that the assumption is potentially invalid for many NLP models. Instead of having an inverted triangle like FNs distribution, we observed a standard triangle like FNs distribution. This is problematic because of the extremely small FNs rate exists across the range from 0 to threshold(e.g., 0.95). In order to have an estimated FNs rate with small margin of error at 95% confidence level, larger order of magnitude data is needed than the expected which is not feasible in production.

Chart, funnel chart

Description automatically generated

Figure 3. The actual FNs distribution illustration.

Figure 4 illustrates the contradiction between the assumed FNs distribution and the actual one. In Figure 4a, the number of FNs with low probability score (e.g., less than 0.5) is negligible. The vast majorities of the FNs is distributed at the high probability score range(e.g., 0.5-T). Meantime, the vast majority of the total sentences is concentrated at the lower probability range(e.g., 0-0.5), and only a small part of the total sentences has large probability score. This allows us to safely ignore the FNs in the range of 0-0.5, and mainly draw samples from the range of 0.5-T. Because its small number of total sentences and large number of FNs, the probability of FNs being sampled is high. Within a certain upper limit of labeling capacity, we can have a good a good estimation of the number of FNs for the current model, which leads to a better recall estimation.

Graphical user interface, application

Description automatically generatedDiagram

Description automatically generated

1. (b)

Figure 4. (a) Stratified assumption illustration. Grey area is not sampleable due to the large amount of total sentences and small amount of positives, whereas the green area is sampleable because of majorities of positives and small amount of total sentences. (b) The actual FNs distribution. No sampleable area exists due to the low FN rates everywhere.

On the contrary, if we have a FNs distribution as depicted in Figure 4b, although large amount of FNs, FNs rate is extreme low at range of 0-0.5 given the much larger order of magnitude of total number of sentences than the FNs number. Meanwhile, low FNs rate in the range of 0.5-T as even smaller number of FNs and smaller number of total sentences despite the number of total sentences is still much larger order of magnitude that the FNs number. As a result, the range from 0-T becomes unsampleable, which means the estimator of FNs rate with small margin of error at 95% confidence level requires much more samples which beyond the labeling capability of annotators.

1. **Results and Contribution**

This project analyzed the FNs distribution of the deployed NLP models’ prediction, which would benefit the recall estimation in the future model monitoring process. The analysis shows potential deficiencies of the current recall estimation approach that could cause misunderstanding the NLP model performance. This work also points out a way to the future recall estimation study for NLP model monitoring purpose, and also show other teams in the organization that the current approach could be biased even for non-NLP models.