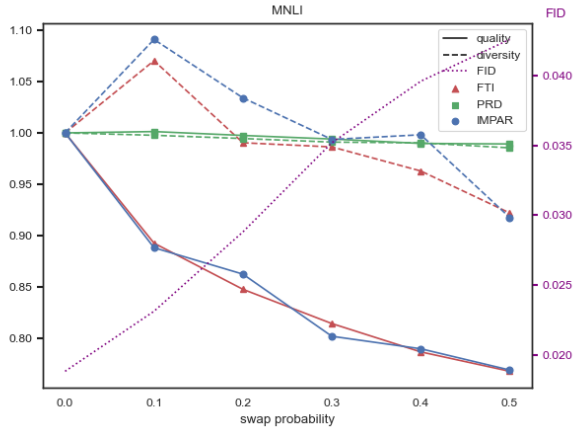

Figure 11. Results for swapping words with a certain probability on MNLI. All metrics are normalized by their respective values obtained on unaltered test sentences, *i.e.* without word swap.

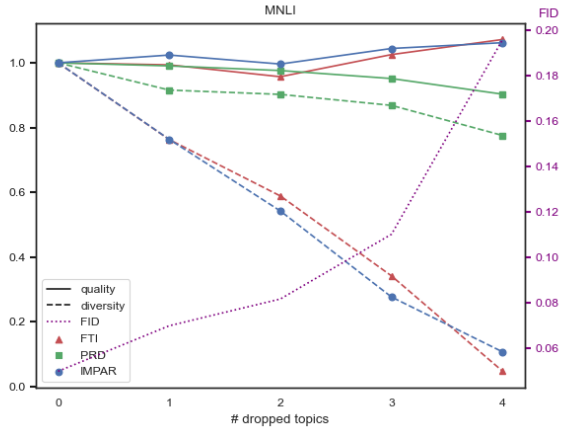

Figure 12. Mode dropping results on MNLI. Metrics are normalized by their respective values on zero dropped topics.

A. Additional Experiments

A.1. Synthetic Experiments on Text Generation

We further tested our approach alongside FID, PRD, and IMPAR on MNLI [33], which assigns 433k sentence-pairs to a specific topic (5 different topics are available in this dataset). Similarly to Semeniuta *et al*. [27], we ignore the pair information and treat each sentence independently. We study each method's sensitivity to the detriment of quality by swapping the words in each test sentence with a certain probability (Section A.1.1). Furthermore, we test sensitivity to diversity detriment by removing sentences of certain topics from the test set (Section A.1.2).

We used the training sentences and testing sentences of MNLI as real and generated samples, respectively. Similarly to Section 4.3, we use the embeddings from USE to obtain a vector representation of each sentence.

A.1.1 Word swap

To test the sensitivity of each method regarding detriment of quality, we swap the words of each generated or test sentence with increasing probability. Hence, as the swapping probability increases, the quality assessment of each method should degrade. On the other hand, diversity should remain somehow constant throughout this experiment. Results are shown in Figure 11.

We observe that FTI and IMPAR behave similarly. Both metrics behave as expected, with their quality assessment deteriorating as the swap probability increases and their diversity remaining somehow consistent over this experiment. Moreover, FID also behaves properly, with its distance increasing with higher swapping probabilities. On the other hand, PRD fails to show any sensitivity to detect word swap.

A.1.2 Mode dropping

We drop sentences from certain topics from our generated or test set, under the intuition that such sentences are representatives of the same data mode. Hence, we try to simulate mode collapse with this synthetic experiment. Since we do not modify any sentence from the test set, we expect the quality assessment to be invariant to the mode drop, while diversity should drop as fewer topics are represented in the test set. Results are shown in Figure 12.

We observe that FTI and IMPAR behave as expected, with their diversity measurement dropping linearly with the drop of topics. FID shows similar behavior. Even though PRD's diversity shows slight signs of detriment, they may not be representative of the abruptness of this experiment, especially at higher degrees of mode dropping.