1. The datasets they used to help their own model as well as features others had come up with for their own system are examples of intrinsic evaluations. An example is them implementing a personal pronoun indicator to see if performance improves, seeing that performance did not improve then taking it away. As well as taking away a feature, such as the intention distribution, checking it against how it was performing, seeing it did better with it, then adding it back the model.

   The extrinsic evaluation is them applying their results for finding hate speech specific tweets. The AI they made and the setup was mostly centered around finding vulgar words and classifying what types vulgarity it is, then applied it to a real world situation such as finding a hate speech tweet that twitter would delete.
2. This is likely because of the data manipulation, with example such as if a token appeared less than 200 times it would be classified as unknown and left out of the n-gram tables. There were many other situations in which they would classify a token an unknown and get counted, but left out of the n-gram tables. It could also be from hyphenated words possibly not being separated, they stated they do try to separate it, but it said they are *usually* separated.
3. False, he states it is not all the unusual

   False, he states it occurs in a number of different writers

   False, he states it occurs once in Shakespeare and once in Marlowe.