There are a lot of different ways one can classify a document, including whether there is some degree of deception in a text. Of course, why people deceive is a multi-faceted issue ...

Phase 1
Consider the following scenarios and discuss how the deceptions involved are similar or different. Consider in particular how much of the deception is in the content and how much is in the form:

1. (Desperate last-minute student) Plagiarism
2. Spam-bots trying to tweet (advertising) content
3. Fan fiction
4. Hotel reviews (anonymously) written by the hotel itself
5. Hotel reviews (anonymously) written by a competing hotel
6. Account hacking
7. Trying to sell a recently uncovered work of Shakespeare’s (which you wrote)
8. Declaring one’s innocence (when guilty)
9. Other scenarios? ...

Phase 2
Consider someone writing a fake review online: what features do you expect to see that differ between a true review and a fake one? Think about specific words, amount of (type of) content, etc.

Phase 3
Now, let’s get out of our groups & experiment ...

1. Let’s identify a restaurant that half of the people in the class like and half don’t.
2. I’m going to quasi-randomly assign you to one of these groups, so that you’re the only one who knows which group you’re in:
   (A) (truth) Like the restaurant, write a review saying you like it
   (B) (lie) Like the restaurant, write a review saying you don’t like it
   (C) (lie) Don’t like the restaurant, write a review saying you like it
   (D) (truth) Don’t like the restaurant, write a review saying you don’t like it

3. Take 5 minutes to write a review.
4. I’ll read the reviews out loud, & you assign each one a category of either True or False.
5. We’ll discuss why we classified as we did, and each writer can discuss how they tried to deceive.

Further reading