Reading 2

Skills:

- Details
- Main idea
- Vocabulary in context
- Understand attitudes

Getting started: How much negative content do you find on Facebook every day?

FACEBOOK IS NOW USING AI TO SORT CONTENT FOR QUICKER MODERATION

Facebook has always made it clear it wants artificial intelligence to handle more moderation duties on its platforms. Today, it announced its latest step toward that goal: putting machine learning in charge of its moderation queue. Here's how moderation works on Facebook. Posts that are thought to violate the company's rules (which includes everything from spam to hate speech and content that "glorifies violence") are **flagged**, either by users or machine learning filters. Some very clear-cut cases are dealt with automatically. In this case, responses could

involve removing a post or blocking an account, for example, while the rest go into a queue for review by human moderators.

Facebook employs about 15,000 of these moderators around the world. Nevertheless, it has been criticized in the past for not giving these workers enough support, employing them in conditions that can lead to trauma. Their job is to sort through flagged posts and make decisions about whether or not they violate the company's various policies.

In the past, moderators reviewed posts more or less chronologically, dealing with them in the order they were reported. Now, Facebook says it wants to make sure the posts that generate more concern are seen first, so it is using machine learning to help. In the future, an amalgam of various machine learning algorithms will be used to sort this queue, prioritizing posts based on three **criteria**: their virality, their severity, and the likelihood they're breaking the rules.

Exactly how these criteria are measured is not clear, but Facebook says the **aim** is to deal with the most damaging posts first. So, the more viral a post is, the quicker it'll be dealt with. The same is true of a post's severity. Facebook says it ranks posts which involve real-world harm as the most important. That could mean content involving terrorism, child exploitation, or self-harm. Posts like spam, meanwhile, which are annoying but not traumatic, are ranked as least important for review. "All content violations will still receive some substantial human review, but we'll be using this system to better prioritize that process," says Ryan Barnes, a product manager with Facebook's community integrity team.

Facebook has shared some details on how its machine learning filters analyze posts. These systems include a model named "WPIE," which stands for "whole post integrity embeddings" and takes what Facebook calls a "holistic" approach to assessing content. This means the algorithms judge various elements in any given post, trying to work out what the image, caption, or poster reveal together.

Facebook's use of AI to moderate its platforms has come in for scrutiny in the past, with critics noting that artificial intelligence lacks a human's capacity to judge the context of a lot of online communication. Especially with topics like misinformation, bullying, and harassment. **Thus**, it can be **nearly** impossible for a computer to know what it's looking at.

Facebook's Chris Palow, a software engineer in the company's interaction integrity team, agrees that AI has its limits but told reporters that the technology could still play a role in removing unwanted content. "The system is about marrying AI and human reviewers to make less total mistakes," says Palow. "The AI is never going to be perfect." When asked what percentage of posts the company's machine learning systems classify incorrectly, Palow didn't give a direct answer but noted that Facebook only lets automated systems work without human supervision when they are as accurate as human reviewers. "The bar for automated action is very high," he said. Nevertheless, Facebook is steadily adding more AI to the moderation mix.

**Adapted from https://www.theverge.com/2020/11/13/21562596/facebook-ai-moderation*

Use of language:

- **Criteria** is a plural word (the singular is criterion). E.g. *What criteria are used for assessing a student's ability*?

Answer the following questions:

1. What is the main idea of the text?
    a. Moderation can only be obtained if humans are checking the platform's content.
    b. AI can help to control undesirable posts on Facebook.
    c. People who violate content policies are always moderated.
    d. Facebook is planning to have a transition to the point that no humans will be needed.

2. The word **flagged** in paragraph 1 is closest in meaning to
    a. deleted
    b. reported
    c. analyzed
    d. destroyed

3. What is the problem Facebook faces with the moderators they employ?
    a. They are not enough.
    b. Their salary is really low.
    c. They lack total support.
    d. Their rights are violated.

4. According to paragraph 3, which are the criteria to review a post soon?
    a. How much it has been shared, the algorithms it uses, and how recent it is.
    b. Its severity, its topic, and the country of origin of the post.

c. How significant the post is, how old it is, and how harmful it can be.
d. Its popularity and how probable it is that it's breaking the policies Facebook has.

5. The word **aim** in paragraph is closest in meaning to
   a. objective
   b. sight
   c. level
   d. meaning

6. What kind of aspects are taken into account to review posts? **Choose two**
   a. The ones that may violate content policies are analyzed first.
   b. They are always checked in chronological order.
   c. The posts that involve spam are removed immediately.
   d. The posts that imply potential violations are reviewed by humans.

7. What is stated about the WPIE model?
   a. It is embedded to every Facebook post.
   b. It tries to analyze the post as a whole.
   c. It judges the visual aspect of the post.
   d. It integrates algorithms from new filters.

8. The word **thus** in paragraph 6 is closest in meaning to
   a. but
   b. since
   c. although
   d. consequently

9. The word **nearly** in paragraph 6 is closest in meaning to
   a. close
   b. very
   c. almost
   d. often

10. What is Chris Palow's attitude towards the use of AI on Facebook?
    a. He says it's almost impossible to implement AI on the platform.
    b. He is positive AI can help them get rid of negative posts.
    c. He agrees AI has problems classifying posts correctly.
    d. He believes one day AI is going to replace humans.

What do you think?

Do you think Facebook should take stronger measures to control the content people post?