Motivation for the analysis:

1) **Analysis 1 – Personal Attacks:** Analyzing the demographic information about the Crowdflower workers is important to understand if their opinions and preferences are representative of the wide population in judgements. It also helps to understand if some of the features from demographics dataset are correlated with how a particular comment is labelled. How toxic a comment is perceived by annotators may depend on both the annotator and the topic. Below are the questions I would like to answer through this analysis:

    - Explore relationships between worker demographics and labeling behavior
    - How consistent are labelling behaviors among workers with different demographic profiles? For example, are female-identified labelers more or less likely to label comments as aggressive than male-identified labelers?
    - If the labelling behaviors are different, what are some possible causes and consequences of this difference?

2) **Analysis 2 – Comparison of all 3 types of hate-speech datasets:** Analyzing the comments datasets of each of the types of hate-speech will help us understand how comments are labelled. How does an annotator decide a particular comment is a personal attack, aggression or toxicity? It throws light on perspective of annotators. Also, it helps understand more detailed nuances of most commonly associated words with each of the three types of hostile speech and how they differ from one another. Below are the questions I would like to answer through this analysis:

    - Analyze the words most commonly associated with each of the three types of hostile speech
    - Are certain words more likely to be associated with comments labelled as hostile speech? Are there certain words that are frequently associated with one type of hostile speech (like "personal attacks") but not others (like "toxicity")?
    - Are these words representative of words that you would associate with hostile speech? Do you think these frequently labelled words are a good representation of hostile speech in online discussions outside of Wikipedia? Of offline discussions? Why or why not?

Presentation of results:

1) **Results of Analysis 1 – Personal Attacks:**

    Analyzing the demographic information about the Crowdflower workers available in the **Personal Attacks** demographics dataset, we can say that these workers represent a very small proportion of population and their opinions and preferences are unlikely to reflect the opinions of the population as a whole. The raw data they are rating may also be insufficient and thus suggest false correlations. How toxic a comment is perceived by annotators may depend on both the annotator and the topic.

    - There are only about 2190 crowdsource workers.
    - The gender distribution for the annotators is as follows - Male: 1349, Female: 840, and other: 1.

    Potential sources of bias:

    - **Features missing values:** Dataframe.count() gives the count of records that have values populated for each column in the dataset. If the dataset has missing values for features for a

large number of examples, then that could be an indicator that certain characteristics of the data set are under-represented. Here we observe that no values are missing for any of the features in comments dataset.

For demographics dataset, we are **missing 1.5%** of age_group feature values, but as this percentage is relatively small, we can ignore it. But if the percentage is very high, we need to understand the implications of missing data and understand why they are missing

- **Omitted variable bias:** The research paper highlights that machines are being trained purely on features extracted only from the comment text instead of including features based on the authors' past behavior and the discussion context. This means that we may be omitting some features which could be directly correlated with the response variable. For example, people's cultural backgrounds and personal sensibilities play a significant role in whether they perceive content as personal attack. So, considering information beyond the text, such as demographic information about the speaker, can improve the accuracy for personal attack detection. A user who is known to write hate speech messages may do so again. A user who is not known to write such messages is unlikely to do so in the future.

  We do not have lot of information in the demographics dataset. Also, without taking the context into account, the models will be not trained to generalize well to unseen examples.

- **Implicit or Experimenter's bias:** For training classifiers, we need to create a corpus that contains a sufficient number and variety of examples of personal attacks. In order to ensure representativeness and overall prevalence of personal attack comments, comments are randomly sampled from the full corpus as well as from the blocked dataset that contains comments made by users who were blocked for violating Wikipedia's policy on personal attacks. But this could be linked to Experimenter's bias because here the experimenter assumes that the comments from blocked dataset are indeed attacks. But thinking deeply, the comments from blocked dataset could also contain biased data. Automation tools could have scored the comments in the blocked dataset and induced some sort bias into the data. Besides this, even if the comments are attacks, they need not be personal attacks. The overall prevalence of personal attacks in the subset of corpus sampled randomly will still be small in the sample data. This means Machine may not have adequate hate speech data to train on.

- **Selection bias:** This type of bias occurs when a model itself influences the generation of data that is used to train it. Blocked dataset contains comments made by users who were blocked for violating Wikipedia's policy on personal attacks, but some of these comments could have been wrongly scored by automation tools based on Machine's learning that has induced biases and are just false positives. Based on the analysis, 65126 comments from the blocked dataset may not be attacks at all in the first place.

- **Unintended bias:** It can be observed that the frequently targeted groups, represented by the identity words such as "black", "muslim", "feminist", "woman", "gay" etc, are over-represented in abusive and toxic comments. This implies the training data used to train machines exhibit the same trend. When the training data used to train machine learning models contain these comments, ML models adopt the biases that exist in these underlying distributions. These identity terms of targeted groups appear far more often in abusive comments. It is much rarer for these words to appear in a positive, affirming statements.

- **False positives:** Flagging identity terms as hate-speech results in False Positives. There is little agreement on what actually constitutes hate speech. Translating an abstract definition into a

clearer and more concrete one can make annotation easier but doing so comes with its own risks. Tools that rely on narrow definitions will miss some of the targeted speech, may be easier to evade, and may be more likely to disproportionately target one or more subtypes of the targeted speech. The general rule that false negatives and false positives should be balanced. However, this assumption ignores the particular stakes of decisions that affect a person's human rights, liberty interests, or access to benefits

Explore relationships between worker demographics and labeling behavior

- For all the gender groups, the age group 18-30 has the highest number of annotators. Females have 395 annotators from age group 18-30. Males have 653 annotators from age group 18-30.

- For all the gender groups, the individuals with bachelor's education constitute major chunk of annotators. Females have 363 annotators holding bachelor's degree. Males have 498 annotators holding bachelor's degree.

- For all the gender groups, the individuals that constitute major chunk of annotators have English not as first language. There are 637 female annotators and 1150 male annotators falling under this category.

How consistent are labelling behaviors among workers with different demographic profiles? For example, are female-identified labelers more or less likely to label comments as personal attack than male-identified labelers?

- We see that females are more likely to call a comment an attack as compared to males. This is because females constitute only 38% and males constitute 61% of crowdflower workers. Some of the other gender groups such as transgenders constitute less than 1%.

- Females and other gender people are underrepresented in the data. This might potentially cause gender and group attribution bias where crowdflower workers belonging to a particular gender might consider comments targeting another gender not an attack.

- There is a tendency to stereotype individual members of a group to which crowdflower workers do not belong. At the same time, there is a preference for the members of the group that crowdflower workers belong to.

- This means males who are overrepresented in the crowdflower group might not consider comments targeting women and other genders as an attack.

- This is one of the many reasons why Females are more likely to label an attack as Personal attack as compared to males. Also, other gender groups are also poorly represented among the crowdflower workers, so we are not able to derive any useful information regarding the other group gender.

If the labelling behaviors are different, what are some possible causes and consequences of this difference?

- Only a small proportion of annotators contributing this content and their opinions and preferences are unlikely to reflect the opinions of the population as a whole. We need to have more participation in human annotation.

- There is little agreement on what actually constitutes hate speech. Getting precise definitions can make annotation easier but doing so comes with its own risks. Machine trained on narrow definitions will miss some of the targeted speech, may be easier to evade, and may be more likely to disproportionately target one or more subtypes of the targeted speech. The general rule that false negatives and false positives should be balanced. However, this assumption ignores the particular stakes of decisions that affect a person's human rights, liberty interests, or access to benefits.

2) **Results of Analysis 2 – Comparison of all 3 types of hate-speech datasets:**

Analyzing the demographic information about the Crowdflower workers available in the **Personal Attacks** demographics dataset, we can say that these workers represent a very small proportion of population and their opinions and preferences are unlikely to reflect the opinions of the population as a whole. The raw data they are rating may also be insufficient and thus suggest false correlations. How toxic a comment is perceived by annotators may depend on both the annotator and the topic.

- There are only about 2190 crowdsource workers.
- The gender distribution for the annotators is as follows - Male: 1349, Female: 840, & other: 1.

Potential sources of bias:

- **Features missing values:** Dataframe.count() gives the count of records that have values populated for each column in the dataset. If the dataset has missing values for features for a large number of examples, then that could be an indicator that certain characteristics of the data set are under-represented. Here we observe that no values are missing for any of the features in comments dataset.

- **Omitted variable bias:** The research paper highlights that machines are being trained purely on features extracted only from the comment text instead of including features based on the authors' past behavior and the discussion context. This means that we may be omitting some features which could be directly correlated with the response variable. For example, people's cultural backgrounds and personal sensibilities play a significant role in whether they perceive content as personal attack. So, considering information beyond the text, such as demographic information about the speaker, can improve the accuracy for personal attack detection. A user who is known to write hate speech messages may do so again. A user who is not known to write such messages is unlikely to do so in the future. We do not have lot of information in the demographics dataset. Also, without taking the context into account, the models will be not trained to generalize well to unseen examples.

- **Implicit or Experimenter's bias:** For training classifiers, we need to create a corpus that contains a sufficient number and variety of examples of personal attacks. In order to ensure representativeness and overall prevalence of personal attack comments, comments are randomly sampled from the full corpus as well as from the blocked dataset that contains comments made by users who were blocked for violating Wikipedia's policy on personal attacks. But this could be linked to Experimenter's bias because here the experimenter assumes that the comments from blocked dataset are indeed attacks. But thinking deeply, the comments from blocked dataset could also contain biased data. Automation tools could have scored the comments in the blocked dataset and induced some sort bias into the data. Besides this, even if the comments are attacks, they need not be personal attacks. The overall prevalence of personal attacks in the subset of corpus sampled randomly will still be small in the sample data. This means Machine may not have adequate hate speech data to train on.

- **Selection bias:** This type of bias occurs when a model itself influences the generation of data that is used to train it. Blocked dataset contains comments made by users who were blocked for violating Wikipedia's policy on personal attacks, but some of these comments could have been wrongly scored by automation tools based on Machine's learning that has induced biases and are just false positives. Based on the analysis, 65126 comments from blocked dataset are not an attack, 64667 comments from blocked dataset are not toxic and 64048 comments from blocked dataset are not aggression.

- **Unintended bias:** It can be observed that the frequently targeted groups, represented by the identity words such as "black", "muslim", "feminist", "woman", "gay" etc, are over-represented in abusive and toxic comments. This implies the training data used to train machines exhibit the same trend. When the training data used to train machine learning models contain these comments, ML models adopt the biases that exist in these underlying distributions. These identity terms of targeted groups appear far more often in abusive comments. It is much rarer for these words to appear in a positive, affirming statements.

- **False positives:** Flagging identity terms as hate-speech results in False Positives. There is little agreement on what actually constitutes hate speech. Translating an abstract definition into a clearer and more concrete one can make annotation easier but doing so comes with its own risks. Tools that rely on narrow definitions will miss some of the targeted speech, may be easier to evade, and may be more likely to disproportionately target one or more subtypes of the targeted speech. The general rule that false negatives and false positives should be balanced. However, this assumption ignores the particular stakes of decisions that affect a person's human rights, liberty interests, or access to benefits

Analyze the words most commonly associated with each of the three types of hostile speech

Frequently occurring top 50 words do not contain hate speech except a couple. This implies that the occurrence of hate-speech words is very less frequent as compared to non-hate speech words. But within the words flagged for hate-speech, Identity terms such as 'deaf', 'blind', 'muslim', 'gay', 'black', 'woman', 'sexuality', 'feminist' appear multiple times. Analysis shows that:

- 717 times identity words appeared in 13590 comments that are perceived as attack
- 812 times identity words appeared in 15362 comments that are perceived toxic
- 712 times identity words appeared in 14782 comments that are perceived aggressive

Are certain words more likely to be associated with comments labelled as hostile speech? Are there certain words that are frequently associated with one type of hostile speech (like "personal attacks") but not others (like "toxicity")?

- The words **'Asian', 'American'** are identity terms associated with personal attacks and toxicity, but not aggression. As **article** points out, there is very low agreement between coders' annotations of text as hate speech. Often, context and minor semantic differences separate hate speech from benign speech. We need clear, consistent definitions of the type of speech to be identified.

- Getting precise definitions can make annotation easier but doing so comes with its own risks. Machine trained on narrow definitions will miss some of the targeted speech, may be easier to evade, and may be more likely to disproportionately target one or more subtypes of the targeted speech.

Are these words representative of words that you would associate with hostile speech? Do you think these frequently labelled words are a good representation of hostile speech in online discussions outside of Wikipedia? Of offline discussions? Why or why not?

- It can be observed that the frequently targeted groups, represented by the identity words such as “black”, “muslim”, “feminist”, “woman”, “gay” etc, are over-represented in abusive and toxic comments. These words are not representative of words that you would associate with hostile speech. This implies the training data used to train machines exhibit the same trend. When the training data used to train machine learning models contain these comments, ML models adopt the biases that exist in these underlying distributions. These identity terms of targeted groups appear far more often in abusive comments. It is much rarer for these words to appear in a positive, affirming statements.

- The frequently labelled words are definitely not a good representation of hostile speech outside Wikipedia. The raw data used for training models is very limited and is not representative of all the words that can be used to express hatefulness. White supremacists have also used innocuous terms, including the names of companies (“Google,” “Skype,” and “Yahoo”) as stand-ins for racial and ethnic slurs according to the article. Users seeking to convey hateful messages could quickly adapt and begin using different novel terms and phrases. Non-English languages are underrepresented and have lower accuracy as they are not well represented on the internet, since the models have fewer examples of those languages to learn from. So, outside Wikipedia which supports very few languages, these words are not a good representation. Also, models have been trained based on Wikipedia corpus which mostly supports informative topics, so outside this domain, the model may not perform well.

Implications for research and product development:

- **What are some other contexts or applications where you would expect the Perspective API to perform particularly well, or particularly poorly? Why?**

  Perspective API may work better in applications where high degree of false positives are acceptable. Some of the identity terms that are overrepresented in toxic comments are flagged for hate-speech. This means we will have more false positives. So, applications such as Author Experience and Author feedback might work really well where the author types in the comment and get a feedback immediately as to how the comment is rated. The impact of this is small and helps the author coin better statements or provide feedback to improve the tool.

  Perspective API will not work in real world applications where the influx of comments or data could be quite different from what the model was trained on. These are the places where high false positives may not be acceptable as it may potentially impact lot of end customers, by blocking them for false alarms.

- **What are some kinds of hostile speech that would be difficult to accurately detect using the approach used to train the Perspective API models?**

When users seeking to convey hateful messages quickly adapt and begin using different novel terms and phrases. If the hateful words are polished and innocuous, the models may not detect it as it is not able to read between the lines and interpret the context and the meaning of those words. Non-English texts and context-based words might not be detected as Machine hasn't trained on those and needs human like interpretation.

- **What are some potential unintended, negative consequences of using the Perspective API for any of these purposes? In your opinion, are these consequences likely or serious enough that you would recommend that the Perspective API not be used in these applications? Why or why not?**

  Perspective API can be used as a firsthand detection tool of hateful speeches, because it can do the validation on a large scale unlike humans. But without human intervention, just relying on the machine's output for vital decision making could be dangerous. In decisions made in the criminal justice or immigration contexts, the question of whether a person is exposed to false-positive or false-negative error could mark the difference between life and death. With more improvisation to the tool based on continuous feedback could definitely improve the accuracy, but monitoring its output and correcting the errors by humans should be an ongoing activity.