Name: Hardik Nahata
Email: nahata.h@northeastern.edu
File: Submission for CS6120:NLP - HW3

ANALYSIS - PART I

VALUE SENSITIVE DESIGN - BACEFOOK INTERNATIONAL

Who is Bacefook International?

- Bacefook International is a global social media network with more than 100K+ users which lets its users post small written posts or witticisms also known as ‘bweets’.
- Bweets can be of maximum length 140 characters.

What do they want to do and why?

- Bacefook International requires us to build an automatic system for tracking the opinions of their users, or in more simple terms, they want a sentiment classification system to understand the opinions of their users.
- They are ready to pay a good sum of money if the applications built are the best in the market.

Who are their users?

- Users of Bacefook International are people who are active social media users who like to share their opinions on ongoing trending topics.

Empirical Investigations

1. Who are the stakeholders involved?

- The social media platform users.
- Bacefook International company.
- The entities, people and companies who are referred to by the platform users in their bweets.

2. What sort of empirical information about the world (psychological, social, cultural, biological, etc.) would be helpful?

- It would be helpful to know the geographical, demographic and social information of the users of the social media platform.
- This information would help us to take precautionary measures of expected bias.

3. What practices or norms or biases are already present? How might these change when technology is introduced? What social and environmental conditions surround materials and production? What aspects of human psychology might we need to understand?

- No data is clean as it is. Since most data is coming from humans, there is always some bias involved.
- Since hotel reviews can be very detailed, especially negative reviews might contain words which inhibit bias on certain groups.
- We need to understand and identify these bias and ensure that the model doesn’t learn modelling this.

Name: Hardik Nahata
Email: nahata.h@northeastern.edu
File: Submission for CS6120:NLP - HW3

Value Investigations

1. What are the various interests and values of the stakeholders? Are these aligned or in tension?

- Bacefook users would want their opinion to be classified correctly by the sentiment classification system. They would want to ensure that their opinion is shown and shared in the right way, as they intend.
- The social media platform - Bacefook International, wants their users to be satisfied by their platform and give good feedback about them. Hence, they would want the sentiment classifier system to work correctly.
- I think the interests of the stakeholders are aligned in terms of the sentiment classification system, as all of them want the system to be unbiased and working efficiently.

2. What harms might result? How should we understand these harms?

- Since the sentiment classification system will use machine learning to train the model, there might be chances of inhibiting bias in the model.
- The harms might be representational or allocative.
- Examples of representational harm in this case could be classifying a bweet which inherently subordinates a group along the lines of identity.
- Example of allocative harm in this case could be classifying a beret that enforces withholding opportunities or resources from a group.

Technical Investigations

1. How might the technology be misused?

- Technology comes with its own challenges. No ML model is completely reliable to avoid bias. The model could pick up some racist comment and classify it as positive if the data it comes across preaches the same. So we need to keep a check.

2. What sort of technical interventions might mitigate or resolve conflict?

Here are some ways to keep the ML model in check:

- Ensure the data fed into the model in clean and unbiased
- Test the model on offensive texts and see that it classifies it negative.

3. What sort of trade-offs might there be between technical optimization and implicated values?

- If we clean the data for bias, we might have to lose on the data samples count, which is not good for the model, as more data is always nice to have.
- The model might need to be retrained as more new data comes in, this would improve performance at the cost of more compute.

Name: Hardik Nahata
Email: nahata.h@northeastern.edu
File: Submission for CS6120:NLP - HW3

ANALYSIS - PART II

1) Characterise the dataset that you used.

1. **Describe the dataset that you used to train your model (how big? label distribution? what is the model being trained on?)**
   - The dataset used for training contains a collection of **hotel reviews**.
   - The data is formatted one review per line
   - The dataset contains **170** total reviews, each having a unique id and assigned label
   - The **labels** denote if the review was **positive (1)** or **negative (0)**
   - The dataset contains **87 positive reviews**
   - The dataset contains **83 negative reviews**
   - The model was trained to do a binary classification on the set of reviews.

2. **Describe the likely effects of training a Naïve Bayes (and Logistic Regression) model on this data and then using it on Bacefook International's data.**
   - Training a Naive Bayes Classifier on the hotel reviews data would probably give us an average performance on the dataset as we ignore new words for classification, and Bacefook's data might be different from the hotel reviews.

   - Training a Logistic Regression model on the hotel reviews data would give us a better performance on the dataset as logistic regression with different features might make the model learn generality in the data and make it perform better on new data.

2) Propose an alternate training set that, given more time, you could use to train your model.

1. **Describe the data that you'd like to use.**
   - Since Bacefook International is a social media platform, the posts and bweets which users might share might not be limited to hotel reviews, hence I would want to consider a dataset which has more diverse samples.
   - We could include political, news, interests, technology and science related samples in the dataset so the model can train on more generic data and in-turn perform better.

2. **Describe the likely effects in terms of performance that using this data would have on your model.**
   - Likely, the model would perform better compared to the hotel reviews dataset.
   - Since the model would have trained on more generic data, it would be able to better classify new samples.

3) Raise any concerns of bias that may be learned by the model from using the data that you proposed in #2

- No data is completely clean. Since most data comes from humans, there might be bias in the train set which the model might pick up.
- We would need to conduct value sensitive design to understand the stakeholders involved

and try to mitigate the bias if any.

4) What are your results?

1. **Document the results of your TextClassifyImproved class on your dev set along (include a table with a breakdown of the effects of the various modifications/features that you tested out)**

Features	Preprocessing	F1 Score (dev)
Bag-of-Words	nltk word tokenizer	0.76
Bag-of-Words Count of positive and negative words	nltk word tokenizer	0.84
Bag-of-Words Count of positive and negative words Length of review	nltk word tokenizer	0.76
Bag-of-Words Count of positive and negative words Logarithm of Length of review	nltk word tokenizer	0.923
Bag-of-Words Count of positive and negative words Logarithm of Length of review Count of first and second person pronouns	nltk word tokenizer	0.96
Bag-of-Words Count of positive and negative words Logarithm of Length of review Count of first and second person pronouns	nltk word tokenizer lowecased text stopwords removed	**1.0**