Homework 2 (Group Homework): Data and Society

This is a group assignment, with your group you work with during recitation. Only one member of your group needs to submit it. If more than one person submits, you will overwrite the previous submission. We will grade only the most recent submission from your group. There are no deadline extensions.

This homework builds on the activity from Week 2 and leads into the Week 3 group presentations. It is structured to encourage each group member to contribute to one of the nine "issue topics" individually, but this does not require that work to be completely apart from the group. The scenarios given can be generalized if desired; the election one can be treated as valid beyond November 5.

Describing the Scenario: Challenger Versus Incumbent

You are a consultant asked to inform on the prospects of a **Challenger** C after C has entered a stage where an **Incumbent** I holds sway. It can be any of:

1. The familiar election scenario—not necessarily for President but any race.
2. C is a new product trying to break into a market where I is the dominant brand.
3. C is a speculative scientific theory that needs popular support for public funding to build labs capable of testing it.

You might be working for C, for I, or for a third party. Your goal is to estimate the chances of C succeeding. Your firm's distinctive concept is to obtain results that are both more precise and more accurate by employing sentiment analysis rather than polling and surveys and voter/consumer profiling as traditionally performed in market research.

Your product has a simple idea:

1. First, you compute an intrinsic sentiment profile S_I of the incumbent official/product/theory from its published literature. That is, you come up with a way to determine the sentiment(s) that *I* represents based on things *I* has written about itself
2. Similarly, you then compute an intrinsic profile S_C for C.
3. Now, your team builds a public sentiment profile S_P of public opinions of relevance to the C-versus-I contest by data-mining social media. *Note: you can assume for your homework that your team just tries to mine any social media you can get your hands on (which can lead to some of the issues we talk about below!)* That is, you use social media to determine what people *want* the winner of *C* or *I* to be.
4. You then use similarity metrics to determine whether S_P is closer to S_I or to S_C, and how much closer.
5. Finally, using this measurement, you compute and output odds of success for C accordingly.

For the election scenario in the current presidential race, the Week 2 activity suggests a simple example with just one profile item and metric. The item is the "heat score" s computed via the Canadian NSERC lexicon. Treating Harris as I and Trump as C, we obtained a substantial gap between s_I and s_C. It makes rough sense that the incumbent party would try to project relative joy and satisfaction with how things are going in their campaign statements and literature, compared to the challenger's party. Since joy is represented by a negative multiplier, we have $s_I \ll s_C$. We may also expect to see public sentiment lie between them: $s_I < s_P < s_C$. The main factor in the election—even more than gaffes or culture—is said to be whether public sentiment on the economy and other major issues aligns more with I or C. So if the measured public "heat score" s_P is closer to s_C than to s_I, we will predict C to win.

In the second scenario, one would organize S_I and S_C along a series of product features and quality issues. The "heat score" implementation might apply to the latter, but otherwise we intend more-refined notions of relevance scores (for matching a social media post to a certain product feature) and sentiment on pertinent aspects. Note that positive consumer sentiment about product I or C should give closeness to S_I and S_C respectively, since we presume a product's own prospectus speaks positively about it. So the basic idea is workable. We'll leave you to imagine analogous ideas in the third scenario.

Finally, before we continue - it should be noted that there are fundamental doubters of your entire enterprise. For instance, at UB you can access the March 2024 final version of a paper by Kellyton Brito, Rogério Luiz Cardoso Silva Filho, and Paulo Jorge Leitão Adeodato (all affiliated to Federal University of Pernambuco in Brazil) titled, "Stop Trying to Predict Elections Only With Twitter." Its subtitle says: "There are other data sources and technical issues to be improved." If you think you can answer this level of doubt, so much the better.

On the other hand, you may find some third-party sources that can help you fix issues—even if the papers are more general than sentiment analysis. An example is a brief white-paper by causalLens titled, "Why Correlation-Based Machine Learning Leads to Bad Predictions." More targeted is this 2021 paper—though it is specific to election prediction.

The Homework Assignment

"So we not only have a killer idea, we have working prototype code that gave (or 'will have given' if you read this right away) plausible results in the week 2 activity. Let's deploy it and let's gooooo..."

Not so fast. Here are nine likely or potential issues. Some of them will have been touched on in the Week 2 lectures, while some others will be on Monday of Week 3. If you find other issues and justify them, that could make a really high-quality creative response.

1. **Domain bias:** The sample taken by the application may be unrepresentative of the population concerned by the C versus I contest.[1]
2. **Skewed proxy:** Sentiment may not be an unbiased estimator of candidate or product or policy choice. Note that this may apply even if the overall sample is representative—so it is not the same as issue 1.
3. **Skewed distribution:** The issue is conveyed by the folk adage that you can drown in a stream whose average depth is only 2 feet. For instance, suppose supporters of C run very hot for C while a larger number of people who will stay with I give middling sentiment for it. The grand average s_P will be pulled close to C by its more-extreme adherents but in a "one-person-one-vote" situation this will mis-predict the outcome.
4. **Flat distribution:** Even if s_P is not skewed, the variance around the mean may be much higher than a model based on the bell curve (or on binomial distribution) would project. This could lead to overconfidence in precision.
5. **Time drift of modeling**: The model training needing to translate the vector(s) of numerical values s_I, s_P, s_C into prediction(s) of the outcome of the C versus I contest may have come from past contests under conditions that no longer hold.
6. **Difficulty validating**: Even if the conditions for a bell-curve distribution are believed to be in effect, there may be insufficient opportunity to verify this. Past trials may be skewed by time drift of important conditions. (So you can lump this

[1] Nate Silver at 3:21pm ET on Sunday 9/15: "This election is basically a competition to see who's more Online. Whichever party wins the competition loses the Electoral College." Meaning not only that the warriors generating the online sentiment and heat are unrepresentative of voters, those voters who hear of it are repulsed toward the other side.

with the previous issue, but here it has more to do with the confidence in the prediction and its range of error rather than generating the prediction itself.)

7. **Mis-classified responses:** The Canadian NSERC lexicon used for the activity does not distinguish different contexts for words. For an example in a possible scenario 3, "inflation" is part of the current incumbent theory of the Big Bang, but it's not only a challenger hot-button in scenario 1, the lexicon was probably scored in that context. (Can you think of other words this caveat may apply to in your chosen scenario?)
8. **Confounding factors:** There may be a factor F that is both a major cause of heightened sentiment and an influence of the C versus I contest that is not in the "profile" of campaign issues or product features, such as to distort the third "directed causal triangle" relation which is our focus.
9. **Selection bias from conditioning:** Suppose we try for increased precision by dividing the population whose sentiments we are mining into segments, and doing our analysis conditioned on each segment. The data under each conditioned run may have biases that are unsuspected because they do not show up in a single large run. [This one will need the Monday Week 3 lecture, but for one aspect that you might be fine with choosing and reading in advance, one can put Simpson's Paradox under this heading. Moreover, various forms of demographic bias—by race or gender or more—may creep in more generally and can be covered as such in your paragraphs.]

These categories do have some overlap, and as stated above, there's also scope for thinking of other issues—more may be added during the term.

Your Task: Submit to UBLearns a presentation that collectively addresses the following:

1. Selects **one scenario** for the C versus I contest
2. Take **four** of these issues listed above. For each issue:

    a. Create 1-2 slides on how that issue could affect the results of a simple deployment of the "s_I, s_P, s_C" idea—perhaps badly enough to make the results invalid. (0.25 points per issue)
    b. Creates 1-2 slides on how your group would improve upon the existing prototype to fix this issue, using insights from class. (0.25 points per issue)
    c. **Note:** For full credit, your answers must be more particular to your chosen scenario than the short descriptions of each issue above.
3. **On your last slide, you should have your Team Participation Statement. Submissions without a correctly formatted statement will receive a 0 for the assignment.** The team participation statement should have the following:
    1. Information on the parts of the assignment that each team member contributed to. We do not need significant details, a few sentences should be enough (e.g. "Persons X and Y contributed to slides 1-4, Person Z helped with slides 1-4 and completed slides 5-7").
    2. A statement by **each team member** that expresses their explicit agreement for the above stated distribution of work. I.e. something like "I agree that this statement accurately reflects the distribution of work in our group. -Your Name".

    Note that we reserve the right to modify individual student grades based on the content of this participation statement.

An End Note

A question that bookends all this is whether current purveyors of already-existing sentiment analysis products are being as scrupulous as we are. On my simple Google search "sentiment analysis for AI products," Google's summarizing AI listed a host of deployed products at the end of its response (which gives three consumer applications that could fold into our scenario 2, and one for managers of multi-family residences). What, if anything, do you find in the documentation of these products (if any)? Beware, however, that sentiment analysis overall is a vast field, and it is easy to get lost in all the

forms of it discussed here by AWS, for instance. We have aimed to focus on one kind of widget and some concrete ideas for its use.