Model Design Specification Document

Analysis of model options explored, the challenges and remedial steps taken

Overview:

Consider a content post d=string()

After tokenization, it appears as-

d=list(T(d))

This tokenized and cleansed text forms the basis of all three approaches explained below. So far 3 Options have been explored-

Option A:

Idea: Measure the occurence of words that attributed to sub-topics of trust. The relative frequnecy of these words in a tweet along with the sentiment state defines the trust score

Approach:

1. Build a sub-topic bag of words(SBOW) for each sub-topic(synonyms, similar words based on human understanding of sub-topic)

2. In T(d), find the % of words that belong to SBOW for each sub-topic S.

3. Aggregate the value generated in step 2 using mean/weighted-mean to obtain trust score.

Challenges:

1. The content of SBOW for each S is relatively sparse(our SBOWs were <30 words)

2. A sparse SBOW leads to very small relative frequency.

3. Even if SBOW size is increased, semantic similarity cannot be associated to a heuristic approach like relative frequency.

4. Response values are also sparse given that the vocabulary/linguistics on social media platforms differs from standard English lexicon(choice of words, abbreviation, tone, level of formality, etc.).

To develop a better understanding of the subject matter being discussed in the sample data, we decided to follow an unsupervised approach to identify discussion topics and the word patterns in these topics.

Option B:

Idea: Using pre-processed content text ct=preprocess(T(d)), establish themes across clusters of topics that are prevalent in the sample data. Based on the term frequencies of repeated terms 'rt' in each topic, build a fuzzy relation between a sub-topic S and rt. The result is an enhanced SBOW upon which relative frequencies and/or semantic similarity can be measured

Approach:

1. Preprocess each content text ct. Add POS tag for each word w in ct

2. Apply LDA using gensim\_models package on all w in all ct in the data set.

3. choose an arbitrary number of topics for clustering.

4. Obtain the top 30(arbitrary) words in each topic.

5. Establish a pattern between the terms with highest relative frequencies to deduce SBOW for each st.

Challenges:

1. The top 30 words for each topic in LDA were generic. Tuning the num\_topics for each LDA did not improve the results

2. While the intent was to identify more trust-relted terms out of the content, the noise levels were still so high that the trust related terms were not reflecting in the unsupervised topics generated by the model. Alternately, we inferred that trust-related or customer-service related talk was too low to be pulled up as a topic on its own.

3. To optimize the topic modeling, we performed all the steps separately for each POS type(noun, adjective, verb) to explore any new patterns in the data. The output still was oriented on generic themes(not trust-related).

Option B:

Idea: Convert word representations of each w in ct into vectors. Aggregate the vector for each ct. Based on labeled scores for Integrity, Transparency and Competency, build an SBOW that represents a theme( +ve or -ve) for each st. The semantic similarity between vectorized ct and SBOW of theme represents the predictor variable for each theme. Based on this representation, fit a regression model to fit a trust score for a given ct.

Approach:

1. Preprocess each content text ct and find word embeddings e for each word w in ct. Embeddings are vectorized representations of w in GloVe.

2. val=sum(e). val is intended to represent a text that is semantically similar to ct. Create a neighborhood of 200 terms(arbitrary) closest to val and store it in vn.

3. Theme th for a ct is known. For every ct belonging to th, find the intersect of all vals. This represents a BOW for th.

4. Thematic score for ct is TS=(Intersection of th BOW and vn)/(word count of ct). The value of each TS for each ct is a predictor variable.

5. Regress upon predictor variables TS and Y=trust score to fit a model

Status:

1. Response variable is available for 300 rows. Out of which ~150 have non-zero scores(trust-related conversations occurring).

2. Step 2 is currently in progress.

Opportunity with option C:

1. Quantifiable values can be generated for each theme.

2. The end goal of finding semantically similar text is a achievable through this approach. Only once the ratios for each theme is computed, we can have more clarity.

Potential gaps:

1. Need to reconfirm the logical interpretation of val.

2. TS in step 4 is heavily dependent on BOW size and content. If the words in BOW for each th have high semantic variance, the overlaps in BOW set and vn will be sparse or potentially disjoint sets.

3. GloVe Embedding computation code takes a large amount of time for each iteration due to high dimensionality of the matrix. Any re-computation significantly slows down the model design process.

Remedial steps:

1. Mitigate the total effort by training the model on smaller sample set.
2. Focus only on one theme/topic and then scale up to incorporate the trust score.
3. Follow modular design for ease of scale and reusablitiy.