Open Internet Sentiment NLP Project

Jocelyn Wang

July 2025

1 Technical Approaches

1.1 Text Similarity (https://huggingface.co/tasks/sentence-similarity)

Sentence similarity models convert input texts into vectors (embeddings) that capture semantic information and calculate how close (similar) they are between them. This task is particularly useful for information retrieval and clustering/grouping.

- Semantic Textual Similarity: Semantic Textual Similarity is the task of evaluating how similar two texts are in terms of meaning. These models take a source sentence and a list of sentences in which we will look for similarities and will return a list of similarity scores.
- Passage ranking: These models take one query and multiple documents and return ranked documents according to the relevancy to the query (query is the "source sentence").

2 Step 1: Identifying Relevance

- SBERT + cosine similarity. Key design decisions are: (1) Choosing representative concept phrases. (2) Setting a reasonable similarity threshold
- Text embedding:
  - "hkunlp/instructor-xl": an instruction-finetuned text embedding model that can generate text embeddings tailored to any task (e.g., classification, retrieval, clustering, text evaluation, etc.) and domains (e.g., science, finance, etc.)
  - HuggingFace Transformers: 'sentence-transformers/all-mpnet-base-v2', 'bert-base-uncased'
  - Sentence Transformers (a.k.a. SBERT; 'all-MiniLM-L6-v2') (https://www.sbert.net/) is the go-to Python module for accessing, using, and training state-of-the-art embedding and reranker models.

3 Step 2: Classify Sentiments

- SBERT + Logistic Regression
- Train (fine-tune) a Custom Classifier using 'google-bert/bert-base-uncased' (require 500 labeled sentences)
- Use a zero-shot model like 'facebook/bart-large-mnli' to classify each sentence into positive, negative, or neutral with no fine-tuning (low accuracy but good for quick test)
- Handling special terms: augment the training data to include labeled examples that feature these terms (DPI; VPN ban)

4 Models Theoretical Comparison and Rationales

- Linear regression / logistic regression: provides a baseline, learns simple patterns
- kNN: interpretable but don't learn pattern, only memorization
- RNN/LSTM: more time-consuming and less powerful compared to transformers; sequential learning, not generalizable to low-data situation since it does not compute similarity
- Text clustering: an enticing choice, but clusters don't imply sentiments; can be difficult to interpret, especially at the boundaries

5 Result - SBERT Embedding + Query-Sentence Ranking

Data source: labeled policy texts from 20110101_Netherlands, 20130101_India.pdf, 20151204_South Africa.pdf, 20161227_China.pdf and some external sources searched online (supported by Gemini)

5.1 Method Rationale

```
# Compute cosine similarity
cos_scores = util.cos_sim(sentence_embeddings, open_internet_query_embeddings)
# shape: (num_sentences, num_queries)

# For each sentence, get the max similarity across all queries
max_scores = cos_scores.max(dim=1).values  # best match per sentence

sorted_indices = max_scores.argsort(descending=True)
for idx in sorted_indices:
    print(f"{max_scores[idx]:.2f} - {sentences[idx]}")
```

5.2 When query data size is small (<20), we encountered misclassification that favors the "closed" side, resulting in "negative" sentiments significantly outnumber "positive" sentiments

(These sentences are originally from the positive sentiment category)

```
Comparing "openness" vs "closedness" scores for each sentence:

[Open: Relevant|0.54] [Closed: Relevant|0.60] [MORE CLOSED] An open Internet is essent
[Open: Irrelevant|0.50] [Closed: Relevant|0.52] [MORE CLOSED] internet providers shall
[Open: Relevant|0.57] [Closed: Relevant|0.60] [MORE CLOSED] It also reaffirms and reco
[Open: Relevant|0.61] [Closed: Relevant|0.50] [MORE OPEN] guaranteeing the freedom of
[Open: Irrelevant|0.47] [Closed: Relevant|0.51] [MORE CLOSED] internet providers shall
[Open: Irrelevant|0.35] [Closed: Irrelevant|0.39] [MORE CLOSED] The FCC was chartered
```

5.3 The Myth of Best Sentence Length/Content

Certain parts of the text are not sentiment-related and having them might misrepresent the text and lead to unintendedly high similarity score. **We are uncertain of the effect of this concern now**.

Example:

*"Manage online activities within the scope of our country's sovereignty according to the Constitution, laws and regulations, protect the security of our country's information infrastructure and information resources, adopt all measures, **including economic, administrative, scientifc, technological, legal, diplomatic and military measures**, to unwaveringly uphold our country's sovereignty in cyberspace. Resolutely oppose all actions to subvert our country's national regime or destroy our country's sovereignty through the network."*

5.4 Problem 5.1 mitigated using data augmentation from ChatGPT

```
Classifying open-internet-opinion sentences:

[Open: Relevant|0.57] [Closed: Relevant|0.51] [MORE OPEN] The Netherlands stanc
[Open: Relevant|0.69] [Closed: Relevant|0.65] [MORE OPEN] Through vigorous and
[Open: Irrelevant|0.47] [Closed: Irrelevant|0.42] [MORE OPEN] The FCC was chart
[Open: Relevant|0.74] [Closed: Relevant|0.61] [MORE OPEN] internet providers sh
[Open: Irrelevant|0.50] [Closed: Irrelevant|0.45] [MORE OPEN] The goal of this
[Open: Relevant|0.65] [Closed: Relevant|0.69] [MORE CLOSED] guaranteeing the fr
[Open: Relevant|0.63] [Closed: Relevant|0.67] [MORE CLOSED] internet providers
[Open: Relevant|0.66] [Closed: Relevant|0.75] [MORE CLOSED] Cyberspace is a glc
[Open: Relevant|0.66] [Closed: Relevant|0.62] [MORE OPEN] It also reaffirms anc
[Open: Relevant|0.55] [Closed: Irrelevant|0.47] [MORE OPEN] Toward this end, th
[Open: Relevant|0.71] [Closed: Relevant|0.53] [MORE OPEN] An appropriate balanc
[Open: Relevant|0.68] [Closed: Relevant|0.58] [MORE OPEN] Privacy, respect for
[Open: Relevant|0.81] [Closed: Relevant|0.66] [MORE OPEN] An open Internet is e
[Open: Relevant|0.67] [Closed: Relevant|0.57] [MORE OPEN] To enhance global coc
[Open: Relevant|0.66] [Closed: Relevant|0.59] [MORE OPEN] World Internet Confer
[Open: Relevant|0.51] [Closed: Irrelevant|0.49] [MORE OPEN] Persist in managing
{'open': 13, 'close': 3}
```

Classifying open-internet sentences: 13 out of 16 correctly identified (81%).

```
Classifying close-internet-opinion sentences:

[Open: Relevant|0.62] [Closed: Relevant|0.73] [MORE CLOSED] No infringement of sovereignty in cyberspace will be tolerated.
[Open: Relevant|0.62] [Closed: Relevant|0.82] [MORE CLOSED] New territories for national sovereignty. Cyberspace has become
[Open: Relevant|0.69] [Closed: Relevant|0.68] [MORE OPEN] Respect for sovereignty in cyberspace, safeguarding cybersecurity
[Open: Relevant|0.50] [Closed: Relevant|0.58] [MORE CLOSED] facilitating the development of an international information se
[Open: Relevant|0.58] [Closed: Relevant|0.71] [MORE CLOSED] Within a state's borders, a state will be controlling its own cy
[Open: Relevant|0.61] [Closed: Relevant|0.72] [MORE CLOSED] Resolutely defending sovereignty in cyberspace.
[Open: Relevant|0.69] [Closed: Relevant|0.77] [MORE CLOSED] Manage online activities within the scope of our country's sove
[Open: Relevant|0.61] [Closed: Relevant|0.59] [MORE OPEN] …guide China's cybersecurity work and safeguard the country's int
[Open: Relevant|0.53] [Closed: Relevant|0.56] [MORE CLOSED] Resolutely oppose all actions to subvert our country's national
[Open: Relevant|0.68] [Closed: Relevant|0.74] [MORE CLOSED] No country should engage in cyber hegemonies, uphold double sta
[Open: Relevant|0.59] [Closed: Relevant|0.58] [MORE OPEN] China will devote itself to safeguarding the nation's interests i
[Open: Relevant|0.64] [Closed: Relevant|0.62] [MORE OPEN] This Law is formulated in order to: ensure cybersecurity; safegua
[Open: Relevant|0.67] [Closed: Relevant|0.65] [MORE OPEN] It promotes that the Internet enriches humanity, and promotes the
[Open: Relevant|0.69] [Closed: Relevant|0.84] [MORE CLOSED] Cyberspace is a new territory for national sovereignty.
[Open: Relevant|0.75] [Closed: Relevant|0.85] [MORE CLOSED] The peoples of all countries are to decide on cyber affairs wit
{'open': 5, 'close': 10}
```

Classifying close-internet sentences: 10 out of 15 correctly identified (67%).

6 Leverage Keywords?

Common Phrases for Each Stance

(blue texts are ones that appear in positive, negative, and neutral stances with varying meaning / significance under different context):

- Open-Internet Common Phrases:
  - Openness & Freedom: open, openness, free, freedom, free flow of information, access, accessibility, transparency, interoperable, common, sharing
  - Human Rights & Democracy: freedom of expression, privacy, democratic, human rights, civil liberties
  - Governance & Cooperation: international cooperation, global governance, shared responsibility, inclusive, collaboration
  - Innovation & Economy: competition, innovation, entrepreneurship, investment, economic development
  - Regulatory Norms: net neutrality, no censorship, device neutrality, rule of law
- Close-Internet Common Phrases:
  - Sovereignty & Control: sovereignty, cyberspace sovereignty, national territory, territorial jurisdiction, control, manage, govern
  - Security & Stability: national security, cybersecurity, information security, public order, regime stability, cyber threats
  - Legal & Regulatory Power: formulate laws, legal measures, constitutional authority, censorship, information management
  - Protectionism & Defense: safeguard, protect, defend, uphold, counter threats, prevent subversion
  - Ideological & Nationalistic Framing: foreign interference, ideological security, online subversion, cultural values, strategic stability
  - Exclusive Framing: within our borders, according to national laws, no foreign interference

Detecting Sentence Relevance: Keywords as Seed Topics + Guided Bert Topic Modeling

- Dataset: 20110101_Netherlands, 20130101_India.pdf, 20151204_South Africa.pdf, 20161227_China.pdf
- Environment note: 'from bertopic import BERTopic' is very sensitive to environmental factors such as the version of packages including transformers, numpy, etc.. I created a new environment to run BERTopic and installed necessary packages (including 'pip install bertopic') as needed to make it run smoothly.
- Problems with topic modeling in identifying sentence relevance:
  - **Even when seed topics (from common phrase keywords) are provided, the topics don't "converge" to what we want (they are still too broad, see sample topics below)**. Potential solution? Experiment with different sets of these keywords. However, it is probably due to the inherent limitation of this (guided) topic modeling —- according to the website, "The topic model will be much more attuned to the categories that were defined previously. However, this does not mean that only topics for these categories will be found. BERTopic is likely to find more specific topics in those you have already defined. This allows you to discover previously unknown topics!"

Seed topics:

```
seed_topic_list = [
# open
["open", "openness", "free", "freedom", "access", "universal"],
["privacy", "democratic", "human rights", "civil", "liberties", "expression"],
# mixed
["international", "cooperation", "global", "shared", "responsibility", "inclusive",
    "collaboration", "borderless", "connectivity", "transparency", "neutrality"],
["laws", "legal", "enforcement", "infrastructure", "standards", "framework",
    "accountability", "governance", "regulation", "compliance"],
# close
["sovereignty", "national", "territory", "foreign", "ideological", "subversion", "cultural", "stability"],
["control", "safeguard", "protect", "prevent", "surveillance", "firewall"]]
```

– **Nevertheless, through the topic name or representative words of each topic Bertopic generates, we do see some reflections of the country's stance, as summarized below:**

* India2013 has both positive & mixed-sentiment labelled sentences ("cooperation", "assessment", "frameworks")

39	-1_systems_research_sectoral_response
34	0_cyberspace_framework_cooperation_cyber space
34	1_practices_products_assessment_critical information infrastructure

* Netherland2011 has mostly positive & neutral sentiment sentences: (see in topic model, words like "international" and "cooperation" is overwhelmingly important).

  While many sentences in the article are labeled as positive, topic modeling reveals a significant focus on potential negative consequences, such as threats and disruptions—suggesting the country's cautious stance and searches for solutions.

  This topic distribution illustrates that the provided keywords may not always align with the resulting topics, reflecting the complexity and variability of the articles. As such, the effectiveness of guided topic modeling in detecting sentence relevance or identifying sentiment warrants investigation.

Count	Name	Representation
119	-1_international_cooperation_security_strategy	['international', 'cooperation',
33	0_cabinet_initiatives_2011_plan	['cabinet', 'initiatives', '2011',
20	1_crime_cyber crime_police_prosecution	['crime', 'cyber crime', 'police'
18	2_netherlands_europe_growth_society	['netherlands', 'europe', 'grow
12	3_cyber security_area cyber security_security_developed	['cyber security', 'area cyber s
10	4_attacks_increasing_cyber attacks_abuse	['attacks', 'increasing', 'cyber

* China2016 has mostly conservative sentences ("sovereignty", "protection", "concerns", "terrorism", "management")

20	-1_using network_political_terrorism_concerns
83	0_international_sovereignty_global_new
28	1_protection_innovation_management_raise

* SouthAfrica2015 has mostly neutral voices ("cooperation", "responsibility", "framework",

"means")

Count	Name
88	-1_responsibility_cooperation_capacity_services
81	0_ncpf_framework_jcps_policy
45	1_means_data_computers_new
18	2_south_africa_south africa_african

7 Use Transformer for Classification? - Lack of Labeled Training Data

In addition to transformer-based embeddings, we can consider these models for classification as well.

Resources:

- `https://www.youtube.com/watch?v=s3LBdmZb00g`: $\approx$ 3000 training data and $\approx$ 2000 test data
  - Dataset Link: `https://www.kaggle.com/crawford/20-newsgroups`
  - Pretrained Models: `https://huggingface.co/transformers/pretrained_models.html)`
- `https://youtu.be/4QHg8Ix8WWQ?si=g0pcQtGj5qLcZXiF,https://youtu.be/8yrD0hR8OY8?si=TO9YxTdQsm23mZCs`: fine-tuning a BERT model for phishing texts classification

Alternatively, we can still use traditional model such as LogisticRegression, but with BERT embeddings.

Later, we can gradually move on to the "neutral" stance and "irrelevant" sentences identification. The definition of such stance & the definition of topic relevance will be key here. To better understand the discourse, we can use t-SNE or UMAP to visualize and discover hidden patterns.