COMPSCI 4NL3: Annotation Setup Report

Group 37

February 26, 2026

1 Team Members

- Abdul-Hadi Siddiqui
- Eduardo Salvacion
- Viransh Shah

2 Dataset Description

For our project, we selected the classic "20 Newsgroups" dataset, a standard in the NLP community for multi-class text classification. This dataset was originally compiled by Ken Lang in the mid-1990s and consists of approximately 18,846 documents. These documents are raw text posts and email-style messages collected from 20 different Usenet newsgroups, covering a wide range of topics from technical discussions to sports and politics.

2.1 Collection and Sourcing

We sourced the data using the `scikit-learn` library's `fetch_20newsgroups()` function. The original data can be found at http://qwone.com/~jason/20Newsgroups/. Our collection process focused on obtaining the clean version of the dataset, where we explicitly opted to remove footers, and quotes. We did this to ensure that our annotators focus on the actual message content rather than relying on metadata shortcuts like newsgroup-specific signatures or quoted text from previous messages. But for our model training, we will also remove headers to further reduce noise as mentioned in the project proposal.

2.2 Format and Structure

Each instance in our dataset represents a single newsgroup post or email message. The dataset is partitioned into a training set of 11,314 documents and a test set of 7,532 documents. In our processed version, each entry corresponds to a block of text associated with one of the 20 topic labels.

The data was collected over a period in the mid-1990s, capturing the early culture of internet forums. We did not perform any additional sampling; we utilized the standard split provided by the source. We did encounter a few documents that were missing text content but had headers or thier headers were gibirish, although the removal of headers occasionally left some posts very short (one or two sentences), which presented an interesting challenge for classification.

2.3 Sensitive Content Disclaimer

Because this dataset consists of unfiltered, user-generated content from public forums in the 1990s, it contains language and views that may be informal, outdated, or controversial. Some newsgroups specifically discuss sensitive topics such as firearms (`talk.politics.guns`), Middle East conflicts (`talk.politics.mideast`), and religious debates. We instructed our team members to be prepared for potentially offensive or sensitive language during the annotation process.

3 Annotation Guidelines and Interface

3.0.1 Overview of the Task

Your job is to read newsgroup posts and assign exactly **ONE** of 20 predefined categories to each. You must determine the primary topic or newsgroup that best represents the content.

3.0.2 Labels and Descriptions

1. **comp.graphics**: Computer graphics, image processing, visualization software, 3D rendering.
2. **comp.os.ms-windows.misc**: Microsoft Windows operating systems, software, and troubleshooting.
3. **comp.sys.ibm.pc.hardware**: IBM PC-compatible hardware, components, and technical issues.
4. **comp.sys.mac.hardware**: Apple Macintosh hardware, peripherals, and technical specs.
5. **comp.windows.x**: The X Window System, Unix/Linux graphical interfaces.
6. **rec.autos**: Automobiles, maintenance, repairs, and car models.
7. **rec.motorcycles**: Motorcycles, riding, maintenance, and motorcycle-related discussions.
8. **rec.sport.baseball**: Baseball games, teams, players, and statistics.
9. **rec.sport.hockey**: Hockey games, teams, players, NHL, and hockey-related discussions.
10. **sci.crypt**: Cryptography, encryption, security algorithms, and data protection.
11. **sci.electronics**: Electronics, circuits, components, and electrical engineering.
12. **sci.med**: Medical topics, diseases, treatments, and healthcare research.
13. **sci.space**: Space exploration, astronomy, NASA, and satellites.
14. **misc.forsale**: Items for sale, selling announcements, and marketplace posts.
15. **talk.politics.guns**: Gun control, firearms legislation, and Second Amendment rights.
16. **talk.politics.mideast**: Middle East politics, conflicts, and international relations.
17. **talk.politics.misc**: General political discussions not covered by other specific categories.
18. **talk.religion.misc**: General religious discussions and interfaith topics.
19. **alt.atheism**: Atheism, secular viewpoints, and critiques of religion.
20. **soc.religion.christian**: Christianity, theology, Bible discussions, and Christian practices.

3.0.3 Rules and Criteria

- **Read Carefully**: Read the entire document before deciding.
- **Primary Focus**: If multiple topics are mentioned, choose the **primary focus**.
- **Specificity**: Choose the most specific category (e.g., Windows graphics drivers go to `comp.os.ms-windows.misc` rather than `comp.graphics`).
- **Political vs. Religious**: If the discussion is about political implications of religion, prioritize the political category (e.g., `talk.politics.mideast` over `talk.religion.misc`).
- **Off-topic**: If the post is completely off-topic or doesn't fit any category, choose the one that is **closest**, but try to avoid forcing it into a category if it clearly doesn't belong.
- **Selling vs. Advice**: If an item is for sale with a price, use `misc.forsale`. If it's asking for advice on what to buy, use the relevant topic category (e.g., `rec.autos`).

3.0.4 Detailed Examples

Example 1: comp.graphics

Text: "Do you have Weitek's address/phone number? I'd like to get some information about this chip."

Label: comp.graphics

Why: The Weitek chip was a graphics coprocessor. This is a technical inquiry specifically about graphics hardware.

Example 2: comp.os.ms-windows.misc

Text: "I have win 3.0 and downloaded several icons and BMP's but I can't figure out how to change the 'wallpaper' or use the icons."

Label: comp.os.ms-windows.misc

Why: This is specifically about Windows 3.0 configuration and OS-level features.

Example 3: comp.sys.ibm.pc.hardware

Text: "SCSI-1 with a SCSI-1 controler chip range is indeed 0-5MB/s... Some PC use this set up too."

Label: comp.sys.ibm.pc.hardware

Why: Discusses technical specs of PC hardware components (SCSI controllers).

Example 4: comp.sys.mac.hardware

Text: "A fair number of brave souls who upgraded their SI clock oscillator have shared their experiences for this poll."

Label: comp.sys.mac.hardware

Why: Discusses hardware modifications specifically for Macintosh computers (SI models).

Example 5: comp.windows.x

Text: "QUESTION: What is the EXACT entry (parameter and syntax please), in the X-Terminal configuration file..."

Label: comp.windows.x

Why: Specifically about X Window System (X11) configuration.

Example 6: rec.autos

Text: "I was wondering if anyone out there could enlighten me on this car I saw... It was called a Bricklin."

Label: rec.autos

Why: Asking for information about a specific car model and its history.

Example 7: rec.motorcycles

Text: "I have a line on a Ducati 900GTS 1978 model... They want 3495, $and I am thinking more like 3$K. Any opinions?"

Label: rec.motorcycles

Why: Seeking advice and opinions on a specific motorcycle model and its pricing.

Example 8: rec.sport.baseball

Text: "Doug Roberts - Ken Hill for NL MVP!! Let's go 'Spos"

Label: rec.sport.baseball

Why: References baseball awards (MVP) and a baseball team (Montreal Expos).

Example 9: rec.sport.hockey

Text: "I think that Mike Foligno was the captain of the Sabres when he got traded to the Leafs."

Label: rec.sport.hockey

Why: Discusses NHL players, teams, and trades.

Example 10: sci.crypt

Text: "Wiretap targets presently use strong encryption, weak encryption, or (the vast majority) no encryption."

Label: sci.crypt

Why: Discusses encryption policy, wiretapping, and cryptographic security.

Example 11: sci.electronics

Text: "I would like to be able to amplify a voltage signal which is output from a thermocouple..."
Label: sci.electronics
Why: Discusses electronic circuit design, amplifiers, and components.

Example 12: sci.med

Text: "There were a few people who responded to my request for info on treatment for astrocytomas..."
Label: sci.med
Why: Discusses a specific medical condition (astrocytomas) and treatments.

Example 13: sci.space

Text: "My understanding is that the 'expected errors' are basically known bugs in the warning system software... before liftoff."
Label: sci.space
Why: Discusses spacecraft launch procedures and systems.

Example 14: misc.forsale

Text: "Reduced Prices! I have a list of things forsale... 1) Black and Decker Duster Plus... Offer: $12."
Label: misc.forsale
Why: Explicit for-sale listing with prices and items.

Example 15: talk.politics.guns

Text: "The term must be rigidly defined in any bill... given this understanding, to consider another class."
Label: talk.politics.guns
Why: Discusses weapons legislation and political debate over gun control.

Example 16: talk.politics.mideast

Text: "January, 1948: Arab Liberation Army attacks Kfar Szold... 14 miles south of Jerusalem."
Label: talk.politics.mideast
Why: Discusses historical and political events in the Middle East region.

Example 17: talk.politics.misc

Text: "Once again, it appears that the one-eyed man has appeared in the land of the sighted and for some strange resaon has appointed himself the ruler..."
Label: talk.politics.misc
Why: General political commentary regarding leadership and power.

Example 18: talk.religion.misc

Text: "the Jews believe that the Covenant between YHWH and the Patriarchs... establishes a Moral Code."
Label: talk.religion.misc
Why: Discusses religious theology and history outside of a purely Christian or atheistic context.

Example 19: alt.atheism

Text: "If you're prepared to say 'Let's round these people up and stick them in a concentration camp without trial'..."
Label: alt.atheism
Why: A moral/philosophical argument about human rights, common in atheism discussions.

Example 20: soc.religion.christian

Text: "Satan has (for some people) loosened the grip on treating the earth as something other than God's int..."
Label: soc.religion.christian
Why: Uses specifically Christian theological terms (Satan, God's intent) to discuss environmentalism.

3.1 Annotation Interface

For our annotation task, we utilized a spreadsheet-based interface. We compiled our data into an Excel workbook (`all_annotations.xlsx`) where each team member had their own dedicated sheet for initial annotations. Separate sheets were created for re-annotations to track overlapping documents.

Annotators were instructed to read the text in the subject and document_text column and enter a label from the 20 predefined categories.

4 Annotation Process and Findings

Each of our three group members spent approximately 1+ hour annotating. We each labeled 100 unique documents, resulting in 300 unique annotations. In the second phase, each member re-annotated 15 documents originally labeled by another teammate, providing us with a total of 45 overlapping documents for agreement calculation. On average, it took approximately 40 seconds to annotate a single instance, though this varied from 30 seconds for obvious posts to several minutes for long, rambling discussions.

Note: All our initial annotations and re-annotations are stored as separate sheets within the `all_annotations.xlsx` file provided with this submission.

4.1 Calculating Agreement

To determine the reliability of our annotation process, we calculated **Krippendorff's Alpha** across the 45 overlapping documents. We chose this metric because it is robust for multiple annotators and we did not annotate every document in the dataset with every annotator.

The calculated value is **0.8590**, where we achieved full agreement on 39, with only 6 documents resulting in disagreements.

4.2 Interesting Data Points and Disagreements

The disagreements we encountered highlight the subjective nature of some topics. Here are three particularly interesting examples:

1. **Document 226 (Original: talk.politics.mideast)**
   Subject: Re: The U.S. Holocaust Memorial Museum: A Costly and Dangerous Mistake
   Text:

```
[DG] THE U.S. HOLOCAUST MEMORIAL MUSEUM: A COSTLY AND DANGEROUS MISTAKE
[DG] by Theodore J. O'Keefe
[DG] HARD BY THE WASHINGTON MONUMENT, within clear view of the Jefferson
[DG] Memorial, an easy stroll down the Mall to the majestic Lincoln Memorial,
[DG] has arisen, on some of the most hallowed territory of the United States of
[DG] America, a costly and dangerous mistake.  On ground where no monument yet
[DG] marks countless sacrifices and unheralded achievements of Americans of all
[DG] races and creeds in the building and defense of this nation, sits today a
[DG] massive and costly edifice, devoted above all to a contentious and false
[DG] version of the ordeal in Europe during World War II, of non-American
[DG] members of a minority, sectarian group.  Now, in the deceptive guise of
[DG] tolerance, the United States Holocaust Memorial Museum begins a propaganda
[DG] campaign, financed through the unwitting largess of the American taxpayer,
[DG] in the interests of Israel and its adherents in America.

[JAKE] After reading the first paragraph, a quick scan confirmed my first
[JAKE] impression: this is a bunch of revisionist and anti-semitic hogwash.

Jake, I'm really disappointed in you. It took you a whole paragraph
to see that it was "bunch of revisionist and anti-semitic hogwash". :-)

The article title "THE U.S. HOLOCAUST MEMORIAL MUSEUM: A COSTLY AND
DANGEROUS MISTAKE" should have been enough! :-)

Tsiel
```

- *Annotator 1*: talk.religion.misc
- *Annotator 2*: talk.politics.misc

Analysis: This post likely touched on the intersection of religious belief and political policy in the Middle East. One annotator focused on the religious arguments being made, while another saw it as a general political debate. This shows how difficult it is to separate "Religion" from "Politics" when the text discusses both simultaneously.

2. **Document 292 (Original: rec.motorcycles)**
Subject: Re: Camping question?
Text:

```
On my LC (RZ to any ex-colonists) I replaced the bolt at the bottom of the barrel
with a tap. When I wanted a coffee I could just rev the engine until boiling
and pour out a cup of hot water.
I used ethylene glycol as antifreeze rather than methanol as it tastes sweeter.

(-:
```

- *Annotator 1*: sci.med
- *Annotator 2*: rec.autos

Analysis: This was a fascinating divergence. The post might have discussed a medical recovery after a vehicle accident. One annotator prioritized the medical discussion (`sci.med`), while another saw the context of vehicles (`rec.autos`). This highlights how a "Science" topic can easily bleed into a "Recreation" topic depending on the reader's focus.

3. **Document 64 (Original: comp.windows.x)**
Subject: Re: Ford and the automobile
Text:

```
: Ford and his automobile.  I need information on whether Ford is
: partially responsible for all of the car accidents and the depletion of
: the ozone layer.  Also, any other additional information will be greatly
: appreciated.  Thanks.
:
SSSSSooooooooooooo!!!!! Its all HIS fault!! Thank God Louis Chevrolet is
innocent! and that guy Diesel, HE otto feel guilty!
```

- *Annotator 1*: comp.graphics
- *Annotator 2*: rec.autos

Analysis: In this case, one annotator likely focused on a technical mention of graphics rendering within an X Windows context, while another might have been confused by a metaphor or a mention of a car in a signature or sidebar. This illustrates the noise in Usenet data.

5 Reflection Questions

(a) What did you learn about the task from doing the annotation? We learned that text classification is far more subjective than it appears on the surface. We initially thought that 20 categories would be distinct enough, but the overlap between "Politics" and "Religion" or "Computer Hardware" and "Operating Systems" is significant. We also realized that without the full thread context, some posts are nearly impossible to classify correctly because they are part of a long-running joke or a very specific niche debate.

(b) What challenges do you expect models to face when learning from your data? Models will likely struggle with the **high vocabulary overlap**. For instance, the five computer-related newsgroups all use words like "driver," "system," "memory," and "performance." Distinguishing between "Mac hardware" and "PC hardware" requires the model to learn very specific brand-name tokens. Additionally, the presence of **sarcasm and informal language** in Usenet posts will likely confuse sentiment-aware models or those that rely on formal grammar.

(c) What surprising things did you observe in the data? We were surprised by how much cross-talk occurs. People in a motorcycle newsgroup might spend half a page arguing about the political implications of helmet laws, which makes the primary topic very hard to define. We also found it interesting to see 1990s-era tech people arguing passionately about 5MB/s SCSI speeds felt like a trip through a digital museum.

(d) Which features do you expect to be useful? Term Frequency-Inverse Document Frequency (TF-IDF) will be essential for identifying keywords that are unique to specific groups (e.g., "NHL" for hockey vs. "RBI" for baseball). **Named Entity Recognition (NER)** could also be powerful for identifying specific OS names, hardware manufacturers, or political figures. We also suspect that **bigrams and trigrams** will be more useful than single words for capturing phrases like "gun control" or "graphics card."

(e) Describe the mental model you came up with to do the task at a high level. Our mental model was hierarchical. We first asked: "Is this post about a hobby, a science, a belief system, or a technical problem?" Once we narrowed it down to a broad group (e.g., "Recreation"), we then looked for "trigger words" that would separate the subcategories (e.g., "puck" $\rightarrow$ hockey, "engine" $\rightarrow$ autos). If the post was purely a debate, we defaulted to the "talk" categories.

(f) Was there anything unclear about the instructions? How would you recommend modifying them to make them more helpful? The instructions were clear for the majority of cases, but "off-topic" posts remained a gray area. We would recommend adding a "None of the Above" or "Off-topic" tag for future iterations. This would allow us to filter out noise from the training set. We would also like to provide a "Priority List", for example, If a post contains both medical advice and a car for sale, prioritize the for-sale category.