COMPSCI 4NL3 - Project Proposal

Winter 2025

Group 37

Team Members

- Abdul-Hadi Siddiqui
- Eduardo Salvacion
- Viransh Shah

Task Title: Multi-Class Text Classification on 20 Newsgroups Dataset

Overview

We are using the “20 Newsgroups dataset” for this experiment, which is used a lot in natural language processing. As stated in the expected dataset size section, this dataset consists of 18846 data samples, where 11314 of them belong to the training set, whilst the rest are associated with the testing set. There are 20 distinct categories/labels associated with the set, which include technology, science, politics, religion, and recreation. It is worth noting that each of the sample points is essentially a raw text post collected from Usenet newsgroups.

The significance of this dataset comes from the fact that it is so large that it supports meaningful machine learning experiments, such as large-scale text classification problems, and it is also ideal for evaluating NLP techniques such as bag-of-words models, TF-IDF, topic modeling, and supervised classifiers.

Some challenges associated with this set include that the sample datapoints are noisy, unstructured, and vary in length (some being much shorter than others). Moreover, regarding the labels for this dataset, some of them have overlapping vocabulary, which means coming up with a distance metric to classify these different categories may be challenging. Preprocessing data may also be challenging due to the presence of informal language or domain-specific terms within the data points.

Task Definition

- Input Data Type: Text documents (newsgroup posts/emails)
- Task Type: Classification

- Number of Classes: 20 classes initially (may reduce to 5-10 subsets if needed)
- Label Type: Single label (each document belongs to exactly one newsgroup)

The 20 Labels:

- comp.graphics
- omp.os.ms-windows.misc
- comp.sys.ibm.pc.hardware
- comp.sys.mac.hardware
- comp.windows.x
- rec.autos
- rec.motorcycles
- rec.sport.baseball
- rec.sport.hockey
- sci.crypt
- sci.electronics
- sci.med
- sci.space
- misc.forsale
- talk.politics.guns
- talk.politics.mideast
- talk.politics.misc
- talk.religion.misc
- alt.atheism
- soc.religion.christian

Data Sources and Collection Plan

We will use the 20 Newsgroups dataset available via scikit-learn and the original source:

- Scikit-learn: https://scikit-learn.org/stable/datasets/real_world.html#newsgroups-dataset
- Original Dataset: http://qwone.com/~jason/20Newsgroups/. Compiled by Ken Lang, containing approximately 20,000 newsgroup documents collected from 20 different Usenet newsgroups in the mid-1990s.
- Dataset Size: 18,846 documents total.
- Features: Text content of newsgroup posts, along with optional metadata (email headers, dates, message IDs).
- Labels: Provided directly in the dataset via newsgroup categories (20 classes).
- Collection Method: Corpus with pre-collected data from Usenet. No scraping is required.

Our Plan:

- Download the dataset directly using scikit-learn's fetch_20newsgroups() function or from the original source URL.
- Use the version without headers/footers/quotes to prevent models from learning metadata-based shortcuts. Preprocess the data further by cleaning URLs and special characters while preserving technical terms, and tokenizing text using standard NLP methods. We will evaluate whether lowercasing, stop word removal, stemming and lemmatization improve model performance.
- Use the entire dataset (18,846 documents) for training, validation, and testing with standard splits.
- If we pivot to a subset of categories, we will select 5-10 related or contrasting categories.
- Re-annotate 300 documents with 1-2 minutes per document (read post, understand context, assign category).
- Total annotation time: Each team member will annotate 100 documents (1.6-3.3 hours per person, 10 hours total for the team).

Expected Dataset Size and Example

For this question, the dataset size we planning to get can be found by using the " fetch_20newsgroups()" from the sklearn.datasets module in python. When extracting information about the "20 Newsgroups", we found:

- **Total: 18,846 documents**
- **Training set: 11,314 documents**
- **Test set: 7,532 documents**
- **Number of labels: 20 topic categories**

Some example datapoints include:

- Sample 1: comp.graphics

  "From: weston@ucssun1.sdsu.edu (weston t)
  Subject: graphical representation of vector-valued functions
  Organization: SDSU Computing Services
  Lines: 13
  NNTP-Posting-Host: ucssun1.sdsu.edu

  gnuplot, etc. make it easy to plot real valued functions of 2 variables
  but I want to plot functions whose values are 2-vectors. I have been
  doing this by plotting arrays of arrows (complete with arrowheads) but
  before going further, I thought I would ask whether someone has already
  done the work. Any pointers??

  thanx in advance

  Tom Weston | USENET: weston@ucssun1.sdsu.edu
  Department of Philosophy | (619) 594-6218 (office)
  San Diego State Univ. | (619) 575-7477 (home)
  San Diego, CA 92182-0303 |
  "

- Sample 2: rec.sport.baseball

  "From: admiral@jhunix.hcf.jhu.edu (Steve C Liu)
  Subject: Re: Bring on the O's

Organization: Homewood Academic Computing, Johns Hopkins University, Baltimore, Md, USA
Lines: 39
Distribution: world
Expires: 5/9/95
NNTP-Posting-Host: jhunix.hcf.jhu.edu
Summary: Root, root, root for the Orioles...

I heard that Eli is selling the team to a group in Cinninati. This would help so that the O's could make some real free agent signings in the offseason. Training Camp reports that everything is pretty positive right now. The backup catcher postion will be a showdown between Tackett and Parent
although I would prefer Parent. #1 Draft Pick Jeff Hammonds may be coming
up faster in the O's hierarchy of the minors faster than expected. Mike Flanagan is trying for another comeback. Big Ben is being defended by coaches saying that while the homers given up were an awful lot, most came
in the beginning of the season and he really improved the second half. This
may be Ben's year.
I feel that while this may not be Mussina's Cy Young year, he will
be able to pitch the entire season without periods of fatigue like last year around August. I really hope Baines can provide the RF support the O's need.
Orsulak was decent but I had hoped that Chito Martinez could learn defense
better and play like he did in '91. The O's right now don't have many left-handed hitters. Anderson proving last year was no fluke and Cal's return
to his averages would be big plusses in a drive for the pennant. The rotation should be Sutcliffe, Mussina, McDonald, Rhodes, ?????. Olson is an
interesting case. Will he strike out the side or load the bases and then get three pop outs? You never know.
The way I see the AL East this year (with personal biases mixed in)
Baltimore
New York

Toronto
Milwaukee
Cleveland
Boston
Detroit
(The top 4 are the only true contenders in my mind. One of these 4 will definitely win the division unless it snows in Hell/Maryland :). I feel that this Baltimore's season to finally put everything together.)

____

|Admiral Steve C. Liu   Internet Address: admiral@jhunix.hcf.jhu.edu|
|"Committee for the Liberation and Intergration of Terrifying Organisms |
|and their Rehabilitation Into Society" from Red Dwarf - "Polymorph" |
|****The Bangles are the greatest female rock band that ever existed!****|
| This sig has been brought to you by... Frungy! The Sport of Kings! |
~~~~~~”

- Sample 3: talk.politics.mideast

“From: hasan@McRCIM.McGill.EDU
Subject: Re: ISLAM BORDERS. (was :Israel: misisipi to ganges)
Originator: hasan@lightning.mcrcim.mcgill.edu
Nntp-Posting-Host: lightning.mcrcim.mcgill.edu
Organization: McGill Research Centre for Intelligent Machines
Lines: 26

In article <4805@bimacs.BITNET>, ehrlich@bimacs.BITNET (Gideon Ehrlich) writes:
|>
|> Hassan and some other seemed not to be a ware that Jews celebrating on
|> these days Thje Passover holliday the holidy of going a way from the
|> Nile.
|> So if one let his imagination freely work it seemed beter to write
|> that the Zionist drean is "from the misisipi to the Nile ".
~~~~~~

the question is by going East or West from the misisipi. on either choice you would loose Palestine or Broklyn, N.Y.

I thought you're gonna say fromn misisipi back to the misisipi !

|> By the way :
|>
|> What are the borders the Islamic world dreams about ??
|>
|> Islamic readers, I am waiting to your honest answer.

Let's say : " let's establish the islamic state first" or "let's free our occupied lands first". And then we can dream about expansion, Mr. Gideon

hasan”