Text Classification in the Wild, Final Project Discussion

CS 490A, Fall 2021

Applications of Natural Language Processing
https://people.cs.umass.edu/~brenocon/cs490a_f21

Brendan O'Connor & Laure Thompson

College of Information & Computer Sciences
University of Massachusetts Amherst

Administrivia

- HW1 grades released
- HW2 final submission due Friday

Text Classification

Input: some text ***x*** (e.g. sentence, document)

Output: a label ***y*** (from some finite label set)

Goal: learn a mapping function ***f*** from ***x*** to ***y***

Classification as reverse engineering

Are labels “true”, “correct”, or “gold standard”?

Classification as reverse engineering

Are labels "true", "correct", or "gold standard"?

The categories / decisions of human annotators might be subjective / arbitrary

Classification as reverse engineering

Are labels "true", "correct", or "gold standard"?

The categories / decisions of human annotators might be subjective / arbitrary

> "Or goal is not to create a system that mimics decisions of a human annotator, but rather to better represent the porous boundaries between labels and identify the [categories] a [text] *could* have been placed…"

Broadwell et al. 2017

The Tell-Tale Hat: Surfacing the Uncertainty in Folklore Classification

Core Question:

- How can classification be used to quantify the variability and uncertainty of folklore indices?

Broadwell et al. 2017

Emic vs. Etic Categories

Emic: from **within** the culture/social group

Etic: from **outside** the culture/social group; cross-cultural

> "…classification is not based upon the structure of the tales themselves so much as the subjective evaluation of the classifier… If a tale involves a stupid ogre and magical object, it is truly an arbitrary decision whether the tale is placed under II A, Tales of Magic (Magic Objects), or II D, Tales of the Stupid Ogre."
>
> –Alan Dunde

Broadwell et al. 2017

Dataset

Folk narratives collected by Danish folklorist, Evald Tang Kristensen, from 1867 to 1924.

31,000+ legends and descriptions of everyday life

36 top-level categories each with multiple secondary categories

>700 secondary categories

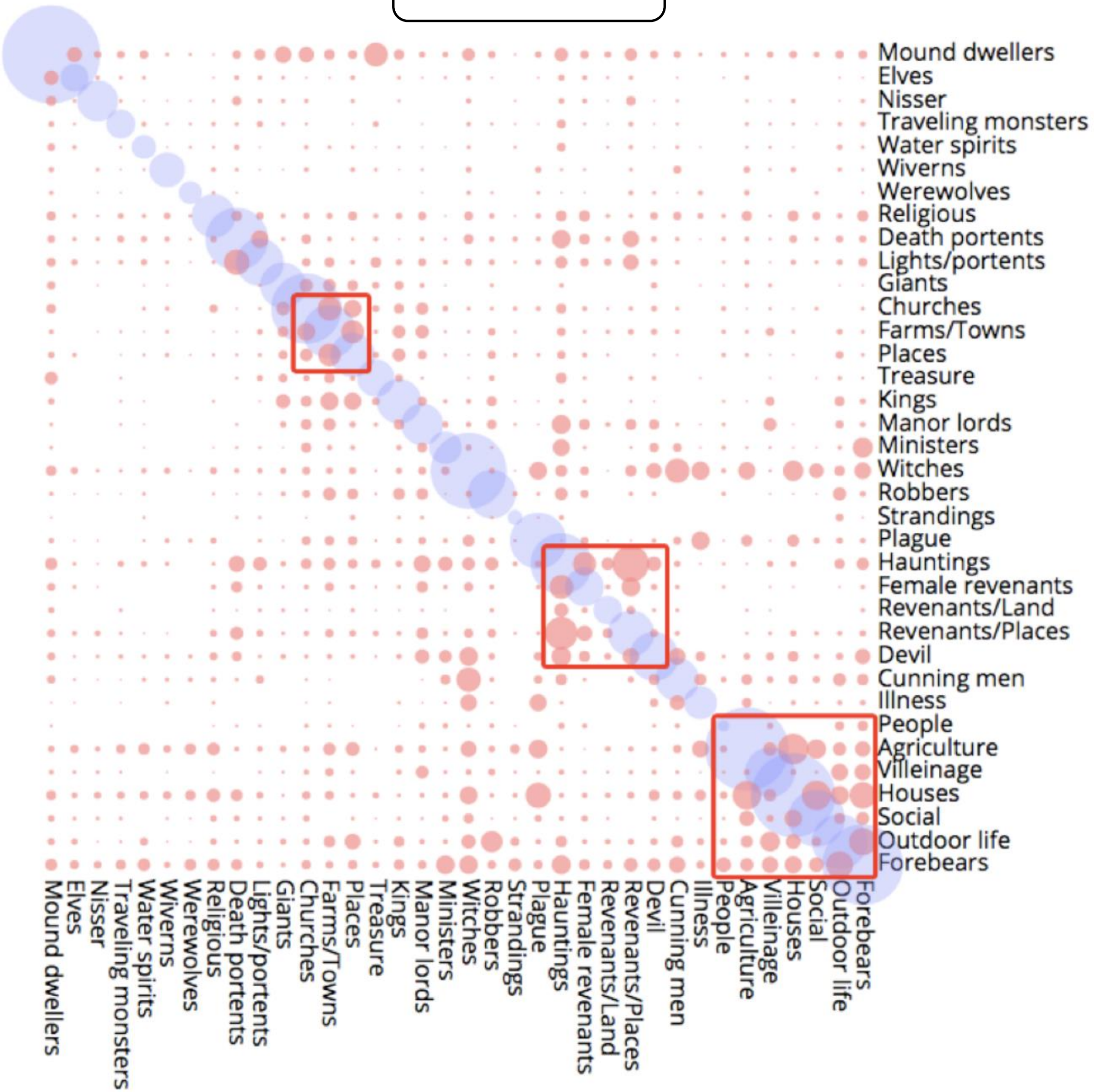
Broadwell et al. 2017

Issues with the classification scheme

- Emic classifications are elided through topic (etic) classification
- Top-level topic categories can be overly broad
  e.g., “Life outdoors”
- Second-level categories can be overly precise
  e.g., “Funeral processions on has seen, or that pass one by” and
  “Funeral processions one has met or followed”

Broadwell et al. 2017

Classifier



Kristensen

Broadwell et al. 2017

Original label	Stories w/orig. label	Stories w/label from NB	Matches	Precision	Recall	F-score
Mound dwellers	2409	2199	1785	0.8117	0.741	0.7747
Elves	255	294	146	0.4966	0.5725	0.5319
Nisser	428	392	312	0.7959	0.729	0.761
Traveling monsters	257	289	159	0.5502	0.6187	0.5824
Water spirits	199	277	111	0.4007	0.5578	0.4664
Wiverns	329	337	234	0.6944	0.7112	0.7027
Werewolves	145	197	98	0.4975	0.6759	0.5731
Religious	693	575	353	0.6139	0.5094	0.5568
Death portents	1113	1127	726	0.6442	0.6523	0.6482
Lights/portents	842	676	418	0.6183	0.4964	0.5507
Giants	545	654	385	0.5887	0.7064	0.6422
Churches	1288	1353	916	0.677	0.7112	0.6937
Farms/Towns	891	1155	530	0.4589	0.5948	0.5181
Places	652	878	360	0.41	0.5521	0.4706
Treasure	411	510	256	0.502	0.6229	0.556
Kings	668	684	366	0.5351	0.5479	0.5414
Manor lords	651	759	323	0.4256	0.4962	0.4582
Ministers	450	460	196	0.4261	0.4356	0.4308
Witches	1830	1771	1074	0.6064	0.5869	0.5965
Robbers	680	754	426	0.565	0.6265	0.5942
Strandings	81	156	40	0.2564	0.4938	0.3375
Plague	893	1018	585	0.5747	0.6551	0.6123
Hauntings	1602	1718	698	0.4063	0.4357	0.4205
Female revenants	607	670	274	0.409	0.4514	0.4292
Revenants/Land	260	302	153	0.5066	0.5885	0.5445
Revenants/Places	897	1086	397	0.3656	0.4426	0.4004
Devil	1030	787	444	0.5642	0.4311	0.4888
Cunning men	753	779	348	0.4467	0.4622	0.4543
Illness	424	494	202	0.4089	0.4764	0.4401
People	125	162	22	0.1358	0.176	0.1533
Agriculture	2154	1763	1294	0.734	0.6007	0.6607
Villeinage	706	789	455	0.5767	0.6445	0.6087
Houses	2363	1888	1314	0.696	0.5561	0.6182
Social	889	994	597	0.6006	0.6715	0.6341
Outdoor life	1287	1143	564	0.4934	0.4382	0.4642
Forebears	2279	1996	1211	0.6067	0.5314	0.5666

Broadwell et al. 2017

Literary Pattern Recognition: Modernism between Close Reading and Machine Learning

Core Question:

What defines the English haiku in the modern period?

Long & So et al. 2016

Is this an English haiku?

Three spirits came to me
And dew me apart
To where the olive boughs
Lay stripped upon the ground;
Pale carnage beneath bright mist.

Long & So et al. 2016

Is this an English haiku?

Three spirits came to me
And dew me apart
To where the olive boughs
Lay stripped upon the ground;
Pale carnage beneath bright mist.

- It's short
- It foregrounds a series of images rather than depict a narrative
- Images are drawn from nature

Long & So et al. 2016

The English haiku as statistical pattern

"This is not […] to reinforce the initial distinction we have made, but to **test its boundaries** and **determine what textual patterns are unique** to each group of texts."

Long & So et al. 2016

Dataset

Haiku – 400 poems

- A translation from a seminal text
- Self-identified as a haiku
  i.e., "haiku" in title
- Identified explicitly as influence by Japanese short verse forms
- 2 categories: translation,
  adaptation

Non-Haiku – 1900+ poems

- Short poems from magazines during the later phases of the haiku's reception
  e.g., *Poetry Magazine*, *Harper's Magazine*, *Lyric West*
- Short: <300 characters

Long & So et al. 2016

Features

Poem as Raw Text

So cold I cannot sleep; and as
I cannot sleep, I'm colder still.

Author Unknown; A 1902 translation by Basil Hall Chamberlain

Poem as a tokenized "bag-of-words"

['so', 'cold', 'i', 'can', 'not', 'sleep', 'and', 'as', 'i', 'can', 'not', 'sleep', 'i'm', 'colder', 'still']

Poem as "bag-of-words" without stopwords (i.e., function words)

['so', 'cold', 'sleep', 'colder', 'still']

Poem as labeled feature set (note that word-order is irrelevant)

[{'cold': True, 'colder': True, 'less_than_20_syl': True, 'sleep': True, 'still': True, 'so': True}, 'haiku']

FIGURE 4. Machine interpretable representations of a single haiku text. Note in the final representation that each feature is assigned a value of "True," indicating its presence in the original text. "Haiku" is the label assigned to the feature vector.

Long & So et al. 2016

Feature Analysis

```
sky = True              not-ha : haiku  =      5.7 : 1.0
shall = True            not-ha : haiku  =      5.0 : 1.0
sea = True              not-ha : haiku  =      5.0 : 1.0
man = True              not-ha : haiku  =      4.3 : 1.0
last = True             not-ha : haiku  =      3.7 : 1.0
snow = True             haiku : not-ha =       3.7 : 1.0
earth = True            not-ha : haiku  =      3.7 : 1.0
blue = True             not-ha : haiku  =      3.7 : 1.0
pass = True             not-ha : haiku  =      3.7 : 1.0
voice = True            haiku : not-ha =       3.7 : 1.0
white = True            not-ha : haiku  =      3.0 : 1.0
house = True            haiku : not-ha =       3.0 : 1.0
```

Long & So et al. 2016

Initial Results

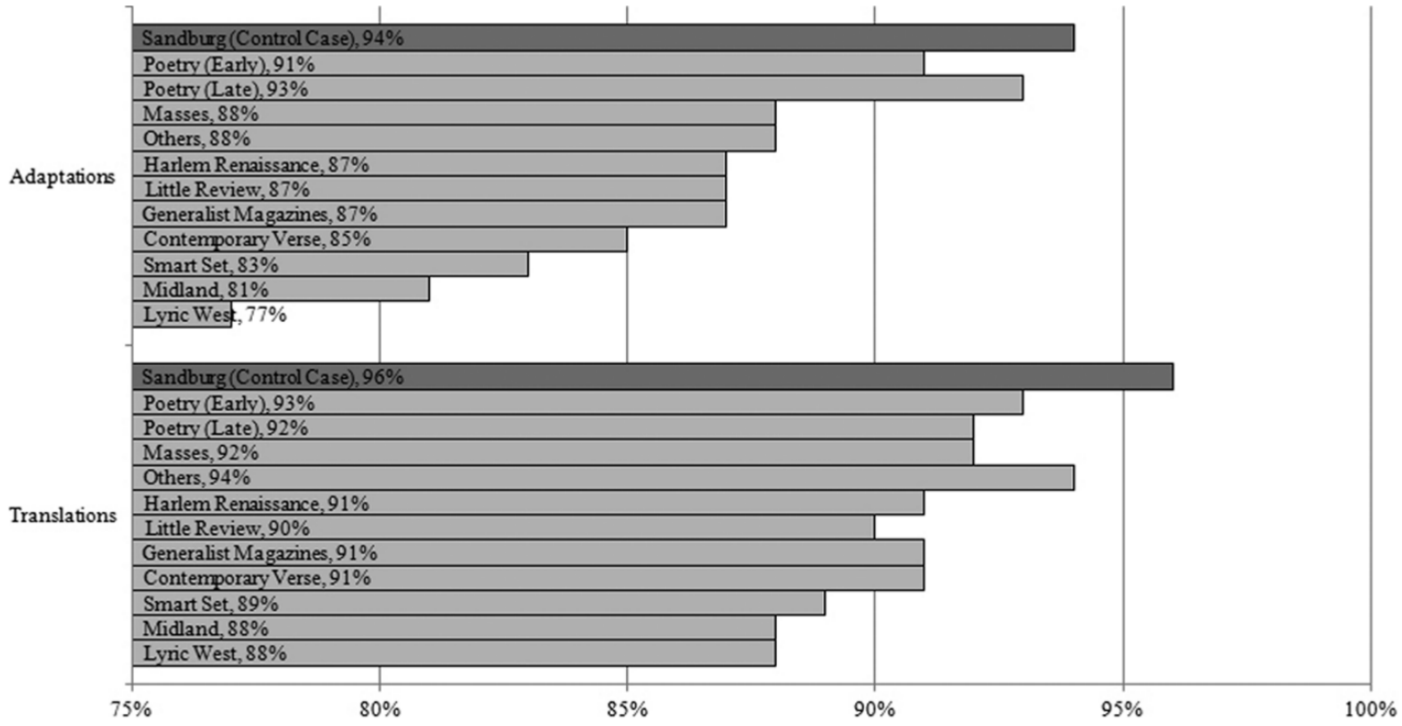


FIGURE 6. Average accuracy scores for one hundred classification tests. The top portion gives the scores for adapted haiku classified against the various short-poem corpora. The bottom portion gives the scores for the translated haiku.

Long & So et al. 2016

After Relaxing Features

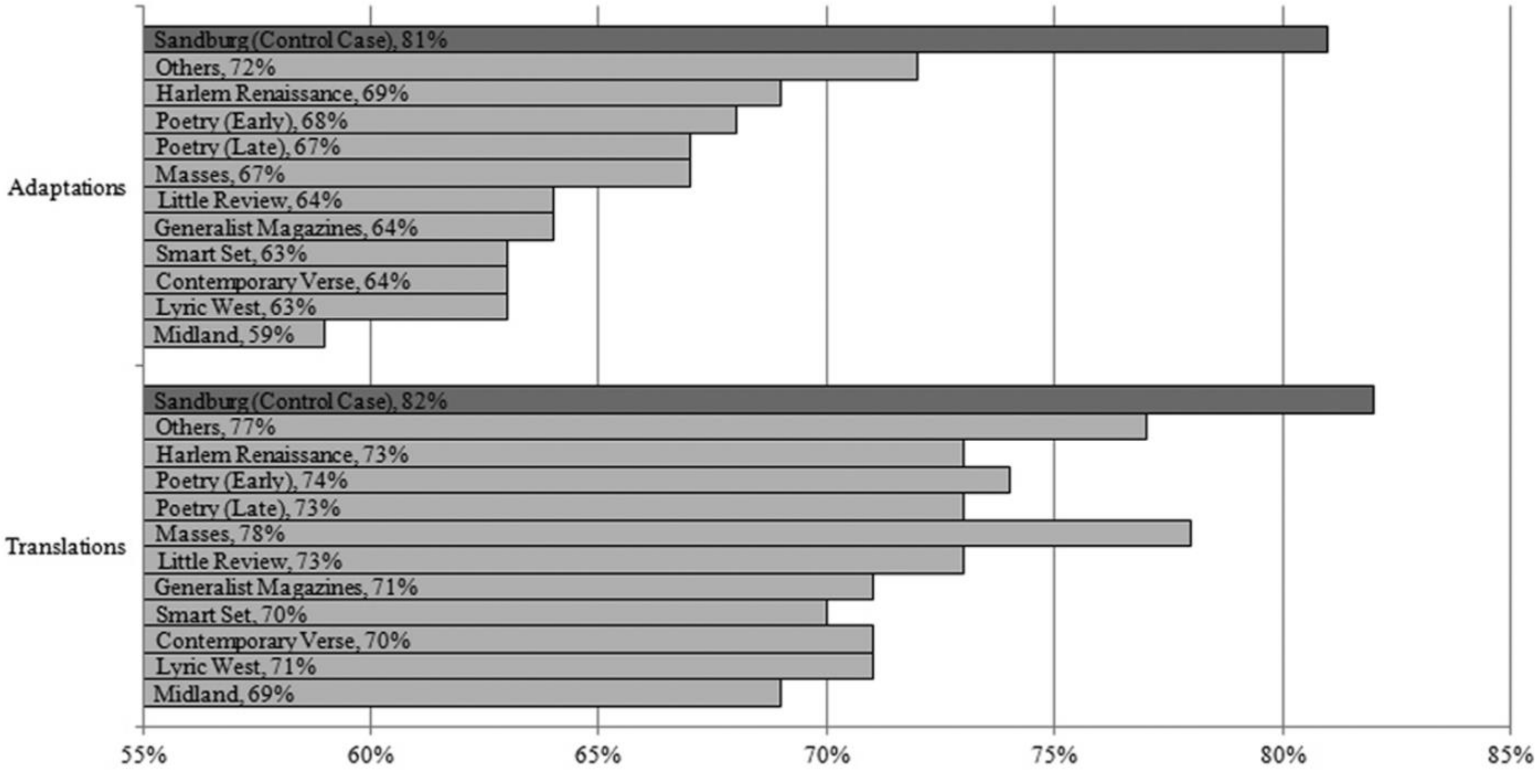


FIGURE 7. Average accuracy scores for one hundred classification tests using a more loosely defined feature set.

Long & So et al. 2016

On Errors

“Rather than correct for the error, what if we consider how it **troubles the initial categorical distinction** built into the procedure? Or better yet, try to generate similar errors so as to **blur the distinction**?”

“What the machine learning literature treats as misclassifications, then, we treat as **opportunities for interpretation**.”

Long & So et al. 2016

Misclassified Poems: Haiku in Waiting

Rain rings break on the pool
And white rain drips from the reeds
Which shake and murmur and bend;
The wind-tossed wistaria falls.

The read-beaked water fowl
Cower beneath the lily leaves;
And a grey bee, stunned by the storm,
Clings to my sleeve.

Long & So et al. 2016

Misclassified Poems: Machine Haiku

When she turns her head sidewise;
The line of her chin and throat
Running down her shoulder
Is as graceful as the undulating motion of the neck of a peacock
Is as smooth as the petals of a Marechel Niel rose.
And her voice
Sounds like a man
Cleaning the rust out of a boiler.

Long & So et al. 2016

Misclassified Poems: In Between

Out of the granite rock I’ve wrested life;
Fending the storm I’ve strengthened root and limb,
Crouching, I hold the plunging chasm’s rim,
As I have braved a thousand years of strife.

Long & So et al. 2016

Final Projects

https://people.cs.umass.edu/~brenocon/cs490a_f21/project.html

Project Overview

Investigate, analyze, and come to research findings about new methods, or insights on previously existing methods.

In groups of 2-4, you will either *build* a natural language processing system or *apply* them to some task.

Your project must: (1) use or develop a dataset, and
(2) report empirical results/analyses with this dataset

Project Components

Proposal: A 2-4 page document outlining the problem, your approach, possible dataset(s) and/or software systems to use.

Progress Report: A 4-8 page document that describes your preliminary work and results

Presentation: An opportunity to present your near-complete project to the class.

Final Report: An 8-12 page document that describes your project and final results.

Project Timeline

- 10/13: Declare project teams
- 10/18: Submit project proposal
- Early Nov.*: Project proposal meeting
- Mid Nov.*: Submit progress report
- Early Dec.*: Class presentations
- 12/16: Submit final report

* = Exact dates to be determined

Where to start

- What *core question(s)* are you trying to answer?
- How will you *operationalize* this question?
- What work are you building off of? What has been done before?
- What experiments will you run?
- How will you measure the success of these experiments?
  e.g., held-out accuracy, error analysis, manual evaluation, etc.

Where to look for related work?

NLP research papers:

- The ACL Anthology is a good place to start
- Some Resources:
  - On how to read research papers
  - On navigating the NLP research space

How to search for papers

- Search keywords in the ACL anthology, Google Scholar, Semantic Scholar
- Look at the papers that a paper references and those that cite it
- Examine other papers by a given author and their lab

Where to look for related work?

A standard web search can also be useful for finding…

- Research blog posts
- Datasets
- Related codebases
- Recorded Talks
- …and more!

Choice of emphasis

- Implementing and developing algorithms and features
- Defining a new linguistic / text analysis task, and tackling it with off-the-shelf NLP software
- Collect and explore a new textual dataset to address research hypotheses about it

A large variety of tasks

Detection Tasks

Classification Tasks

Prediction Tasks

- Predict external information from text (e.g. movie revenue, post popularity, stock volatility, etc.)

Structured Linguistic Prediction

- Relation, event extraction
- Narrative chain extraction
- Parsing

Text Generation Tasks

- Machine Translation
- Summarization & Normalization
- Poetry / Lyric generation

End-to-End Systems

- Question Answering
- Conversational dialogue systems

Visualization & Exploration

- Temporal analysis of events
- Topic modeling & clustering

For more dataset and task ideas

- Look at resources listed in 9/28 lecture slides
- Shared task websites
  - SemEval: Series of semantic evaluation tasks
    - SemEval 2022 tasks (look at older ones, for access to data)
    - SemEval 2021 tasks
  - CoNLL shared tasks

Some projects from last year

Text Classification

- Song genre classification using lyrics
- Comparing models for multi-labeled classification of book genres
- Distinguishing between 19^{th} and 20^{th} century literature
- Predicting political slant in news comments
- Classification of political views on Reddit
- Classifying BBC news articles into their section/category types
- Language classification

Some projects from last year

Detection Tasks

- Paraphrase detection
- Toxicity level detection in social media posts

Prediction Tasks

- Estimating stock volatility from news articles
- r/AmITheAsshole verdict prediction
- Predicting tweet popularity

Text Generation Tasks

- Text summarization for lectures

End-to-End Systems

- FAQ answering
- Medical diagnosis chatbot

Visualization & Exploration

- Sentiment analysis of songs throughout time
- Sentiment analysis of r/wallstreetbets

Exercise / In-Class Activity

JEOPARDY!

Category Analysis

Evan Risas & Alisa Kotliarova

Task: Analyze each question-answer pair to determine which broad category it most closely fits, then predict category frequency for future Jeopardy games.

Dataset
200,000+ Jeopardy! Question & Answer pairs

→

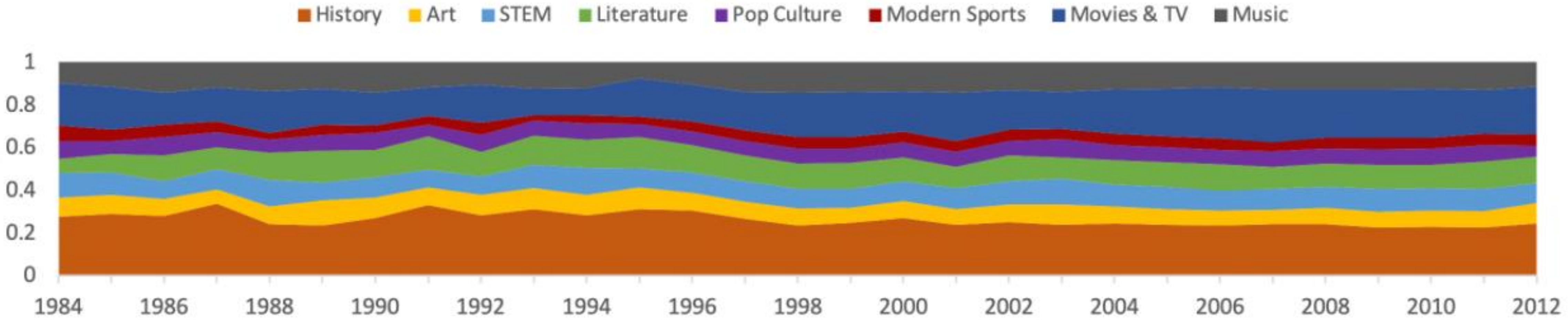
Approach
Classify into custom categories using NLP model built on Word2Vec

→

Observe
Examine category popularity over time

→

Predict
Predict future category frequency



Brainstorming Session

Having trouble finding a group?

..checkout Piazza.

The Search for Teammates feature is coming soon!