Text Classification in the Wild

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 have been released
- HW2 Annotated dataset due tonight!
- Final submission due Friday

Text Classification

Input: some text $\boldsymbol{x}$ (e.g. sentence, document)

Output: a label $\boldsymbol{y}$ (from some finite label set)

Goal: learn a mapping function $\boldsymbol{f}$ from $\boldsymbol{x}$ to $\boldsymbol{y}$

Thumbs up?

Sentiment Classification using Machine Learning Techniques

Core Question:

Can machine learning techniques be used to classify documents by overall sentiment?

x: movie reviews

y: + / - label

Pang et al. 2002

Why might this be a hard task?

Pang et al. 2002

Why might this be a hard task?

“How could anyone sit through this movie?”

Pang et al. 2002

Why might this be a hard task?

“This film *should* be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is *attempting* to deliver a good performance. However, it *can’t* hold up.”

“I *hate* the Spice Girls. ...[3 things the author hates about them]... Why I saw this movie is a really, really, really long story, but I did, and one would think I’d *despise* every minute of it. But... Okay, I’m really *ashamed* of it, but I enjoyed it. I mean, I admit it’s a *really awful* movie ...the ninth floor of *hell*...The plot is such a *mess* that it’s *terrible*. But I loved it.”

Pang et al. 2002

Dataset

IMDB reviews: 700 positive (+), 700 negative (−)

Available at: cs.cornell.edu/people/pabo/movie-review-data/

Labels:

- Extracted from review text
- Label strongly positive reviews as +
- Label strongly negative reviews as −
- Others considered neutral and discarded

Pang et al. 2002

Dataset

Data Curation:

- For each author, only include at most 20 + and 20 – reviews
- Extract text from html
- Remove explicit ratings (“*** out of *****”) & boilerplate text
- Treat punctuation as individual tokens
- Lowercase text

```
these are words that could be used to describe the
emotions of john sayles' characters in his latest
, limbo . but no , i use them to describe myself
after sitting through his latest little exercise
in indie egomania . i can forgive many things .
but using some hackneyed , whacked-out , screwed-
up * non * -ending on a movie is unforgivable . i
walked a half-mile in the rain and sat through two
hours of typical , plodding sayles melodrama to
get cheated by a complete and total copout finale
. does sayles think he's roger corman ?
```

Pang et al. 2002

Even preexisting datasets can be messy

filmcritic . com presents a review from staff member james brundage . you can find the review with full credits at http : //filmcritic . com/misc/emporium . nsf/2a460f93626cd4678625624c007f2b46/c97ebb11df0b98398825694f005571d7 ? opendocument he is duncan macleod of the clan macleod . he's been pimpin' it since he was born in the village of glennfillan in 15somethingsomething , and he continues to pimp it in modern day . he is immortal and he cannot die .

all of my film reviews are archived at http : //us . imdb . com/m/reviews_by ? justin + felix this review has been submitted to the shrubbery http : //www . theshrubbery . com any comments about this review ? e-mail me at justinfelix@yahoo . comscreen story by kevin yagher and andrew kevin walker . inspired by the short story the legend of sleepy hollow by washington irving .

1939 , g , 222 minutes [3 hours , 42 minutes] starring : viven leigh (katherine scarlett o'hara-hamilton-kennedy-butler) , clark gable (captain rhett butler) , olivia de havilland (melanie wilkes) , leslie howard (ashley wilkes) ; written by sidney howard ; produced by david o . selznik ; directed by victor fleming ; based on the novel by margaret mitchell . seen july 8 , 1998 at the crossgates cinema 18 , (albany , ny) , theater #7 , at 8 : 15 p . m . with my mom using hoyts cinema cash . **[theater rating : * * * 1/2 : very good sound , picture , and seats]**

Pang et al. 2002

Even preexisting datasets can be messy

–

```
filmcritic . com presents a review from staff member james brundage . you can find the
review with full credits at http : //filmcritic . com/misc/emporium .
nsf/2a460f93626cd4678625624c007f2b46/c97ebb11df0b98398825694f005571d7 ? opendocument he
is duncan macleod of the clan macleod . he's been pimpin' it since he was born in the
village of glennfillan in 15somethingsomething , and he continues to pimp it in modern
day . he is immortal and he cannot die .
```

+

```
all of my film reviews are archived at http : //us . imdb . com/m/reviews_by ? justin + felix
this review has been submitted to the shrubbery http : //www . theshrubbery . com any comments
about this review ? e-mail me at justinfelix@yahoo . comscreen story by kevin yagher and andrew
kevin walker . inspired by the short story the legend of sleepy hollow by washington irving .
```

+

```
1939 , g , 222 minutes [3 hours , 42 minutes] starring : viven leigh ( katherine scarlett o'hara-
hamilton-kennedy-butler ) , clark gable ( captain rhett butler ) , olivia de havilland ( melanie
wilkes ) , leslie howard ( ashley wilkes ) ; written by sidney howard ; produced by david o .
selznik ; directed by victor fleming ; based on the novel by margaret mitchell . seen july 8 ,
1998 at the crossgates cinema 18 , ( albany , ny ) , theater #7 , at 8 : 15 p . m . with my mom
using hoyts cinema cash . [theater rating : * * * 1/2 : very good sound , picture , and seats]
```

Pang et al. 2002

Word List Baselines

Baseline	Proposed word lists	Accuracy	Ties
Human 1	+: dazzling, brilliant, phenomenal, excellent, fantastic −: suck, terrible, awful, unwatchable, hideous	58%	75%
Human 2	+: gripping, mesmerizing, riveting, spectacular, cool, awesome, thrilling, badass, excellent, moving, exciting −: bad, cliched, sucks, boring, stupid, slow	64%	39%

Pang et al. 2002

Word List Baselines

Baseline	Proposed word lists	Accuracy	Ties
Human 1	+: dazzling, brilliant, phenomenal, excellent, fantastic −: suck, terrible, awful, unwatchable, hideous	58%	75%
Human 2	+: gripping, mesmerizing, riveting, spectacular, cool, awesome, thrilling, badass, excellent, moving, exciting −: bad, cliched, sucks, boring, stupid, slow	64%	39%
Human 3 + stats	+: love, wonderful, best, great, superb, still, beautiful −: bad, worst, stupid, waste, boring, ?, !	**69%**	16%

5%

How many features 14

Freq.

Pang et al. 2002

Results

Presence (binary) is good

Naive Bayes

Logistic Regression

	Features	# of features	frequency or presence?	NB	ME	SVM
(1)	unigrams	16165	freq.	**78.7**	N/A	72.8
(2)	unigrams	"	pres.	81.0	80.4	**82.9**
(3)	unigrams+bigrams	32330	pres.	80.6	80.8	**82.7**
(4)	bigrams	16165	pres.	77.3	**77.4**	77.1
(5)	unigrams+POS	16695	pres.	81.5	80.4	**81.9**
(6)	adjectives	2633	pres.	77.0	**77.7**	75.1
(7)	top 2633 unigrams	2633	pres.	80.3	81.0	**81.4**
(8)	unigrams+position	22430	pres.	81.0	80.1	**81.6**

* binary features!

love_verb
love_proper noun

10K aren't so important

Figure 3: Average three-fold cross-validation accuracies, in percent. Boldface: best performance for a given setting (row). Recall that our baseline results ranged from 50% to 69%.

Pang et al. 2002

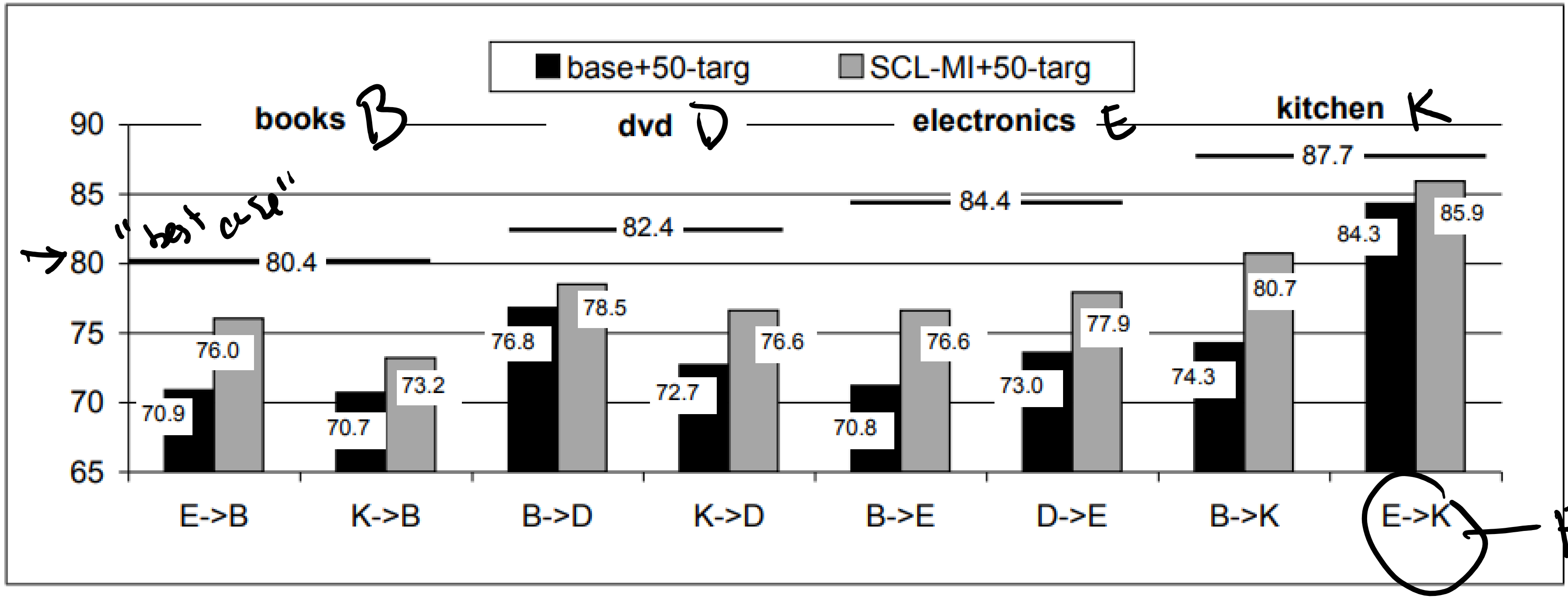
Sentiment expression varies across domains

domain\polarity	negative	positive
books	plot <num>_pages predictable reading_this page_<num>	reader grisham engaging must_read fascinating
kitchen	the_plastic poorly_designed leaking awkward_to defective	excellent_product espresso are_perfect years_now a_breeze

Table 2: Correspondences discovered by SCL for books and kitchen appliances. The top row shows features that only appear in books and the bottom features that only appear in kitchen appliances. The left and right columns show negative and positive features in correspondence, respectively.

Blitzer et al. 2007

Sentiment expression varies across domains



Fairly similar

Figure 2: Accuracy results for domain adaptation with 50 labeled target domain instances.

Blitzer et al. 2007

On Positivity Bias in Negative Reviews

Core Question:

How are positive words used within negative reviews?

> "Food was ok…***not*** the money they charge. I was *unimpressed* will ***not*** return. I was excited to try this place and was so *disappointed* as my expectations were high. Service ***not*** great and the parking is *awful*."

Aithal & Tan et al. 2021

Dataset Analysis

Negative reviews have more positive words than negative words

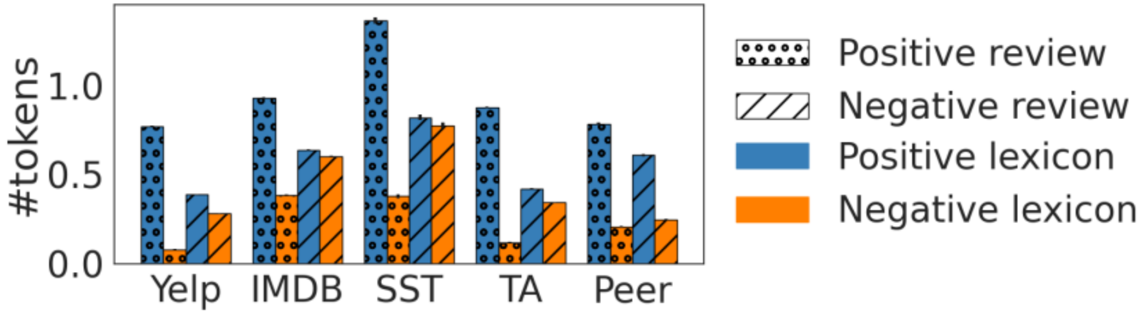


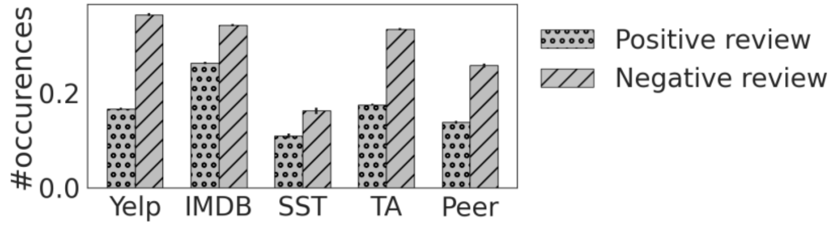
Figure 2: Number of positive and negative words based on Vader. Negative reviews have more positive words than negative words.

Aithal & Tan et al. 2021

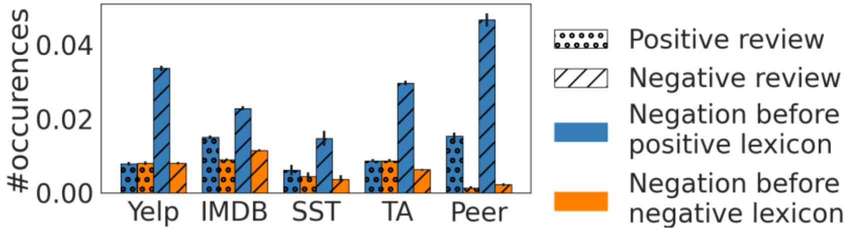
Dataset Analysis

Dataset	Positive words associated with negations
Yelp	recommend, sure, like, good, care, great, special, impressed, fresh, help, ready, enjoy, friendly, honor, helpful, clean, happy, accept, greeted, amazing
IMDB	like, care, funny, help, sure, recommend, good, save, fit, great, special, interesting, enjoy, well, play, better, giving, original, convincing, true
PeerRead	clear, sure, convincing, convinced, ready, well, true, clearly, surprising, novel, convincingly, recommend, guarantee, improve, interesting, support, satisfactory, help, acceptable, convince

Table 2: Most frequent positive words that immediately follow negations in negative reviews, based on Vader.



(a) Overall negation.

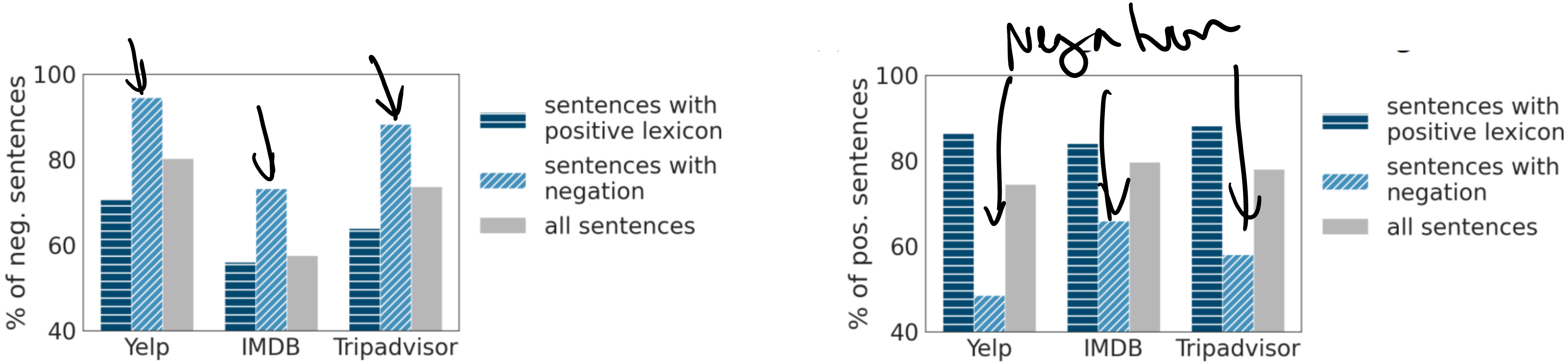


(b) Negation before positive and negative lexicons.

Figure 3: Negative reviews generally have more negations at the sentence level (Figure 3a). Among those negations, Figure 3b shows that there are substantially more negations before positive lexicons in negative reviews than any other combinations.

Aithal & Tan et al. 2021

Sentence analysis using classification



(a) Fractions of negative sentences in negative reviews. (b) Fractions of positive sentences in positive reviews.

Figure 4: Sentence-level prediction results based on fine-tuned BERT classifiers. In negative reviews, sentences with positive words tend to be negative, and sentences with negations are overwhelmingly negative. In comparison, sentences with negations are more balanced (44.7% negative) in positive reviews.

<u>Aithal & Tan et al. 2021</u>

#SupportTheCause: Identifying Motivations to Participate in Online Health Campaigns

Core Question:

How does participant motivation *impact* the amount of campaign donations raised?

#s

classification

Nguyen et al. 2015

Dataset

Movember Profiles

- US & UK Movember participants
- Collected May 2015
- 166,222 US and 138,546 UK profiles

Twitter Data

- Link Movember participants with Twitter accounts using tweets that link to a Movember profile in 2013 or 2014
- Match 5,519 users
- Collect tweets from 10/18 to 12/14 (two weeks before & after campaign)

Nguyen et al. 2015

Annotation

Labels based on the Social Identity Model of Collective Action:

- Injustice
  'my dad', 'I had testicular cancer', 'b/c men's health is important to me'
- Social Identity
  'my friends asked me to join', 'a great excuse to grow a stache'
- Collective Efficacy
  'this campaign can make a difference!'

Participants can have multiple motivations

	Train	Test
# Participants	1,494	614
% US / UK	54.8/45.2	53.3/46.7
% Injustice	37.6	40.2
% Social identity	48.7	46.9
% Collective efficacy	36.1	35.0

Table 2: Dataset statistics

Nguyen et al. 2015

Results

Classifiers trained on Movember profiles are fairly accurate

precision recall F1

Features	Injustice				Social Identity				Collective Efficacy			
	P	**R**	**F_1**	**AUC**	**P**	**R**	**F_1**	**AUC**	**P**	**R**	**F_1**	**AUC**
Tokens	0.813	0.789	0.801	0.833	0.768	0.792	0.779	0.790	0.595	0.656	0.624	0.708
LDA	0.789	0.802	0.795	0.829	**0.809**	0.795	**0.802**	**0.815**	0.514	**0.688**	0.588	0.669
Length	0.644	0.615	0.629	0.693	0.526	0.632	0.574	0.564	0.419	0.642	0.507	0.582
Country	0.422	0.559	0.481	0.522	0.495	0.493	0.494	0.524	0.373	0.498	0.426	0.523
All	**0.823**	**0.810**	**0.816**	**0.846**	0.777	**0.799**	0.788	0.798	**0.597**	0.660	**0.627**	**0.710**

Table 1: Results free-text motivations: precision (P), recall (R), F_1 score and AUC.

Nguyen et al. 2015

Results

It's difficult to predict motivation from tweets

Features	Injustice				Social Identity				Collective Efficacy			
	P	R	F_1	AUC	P	R	F_1	AUC	P	R	F_1	AUC
1: Tokens	0.456	0.445	0.451	0.544	**0.528**	0.563	0.545	**0.559**	**0.394**	0.465	0.426	**0.540**
2: URLs	0.421	0.304	0.353	0.511	0.469	**0.736**	**0.573**	0.500	0.360	0.209	0.265	0.504
3: Mentions	0.435	0.340	0.382	0.522	0.477	0.694	0.566	0.511	0.360	**0.721**	**0.480**	0.515
4: Effort	0.434	0.518	**0.472**	0.532	0.489	0.531	0.509	0.520	0.363	0.498	0.420	0.513
5: LDA	0.427	0.510	0.465	0.525	0.512	0.538	0.525	0.542	0.378	0.521	0.438	0.530
6: Behavior	0.415	**0.526**	0.464	0.514	0.463	0.410	0.435	0.495	0.360	0.581	0.445	0.513
1+3+4+5+cntry	**0.463**	0.453	0.458	**0.550**	0.520	0.542	0.531	0.550	0.381	0.419	0.399	0.526

Table 3: Results on tweets: precision (P), recall (R), F_1 score and AUC.

Nguyen et al. 2015

Motivations & Campaign Behavior

low higher

	% Injustice	% Identity	% Efficacy
UK	31.0	49.7	46.1
US	37.6	50.3	32.1

Table 5: Motivation distribution based on automatic annotation (n=90,484). Note that participants may have multiple motivations.

tend to raise the most low

	Injustice	Identity	Efficacy
UK ($)	203.74	128.36	123.39
US ($)	234.47	156.07	169.03

Table 6: Average amount raised (n=90,484). British pounds were converted in dollars following the exchange rate in November 2013.

Nguyen et al. 2015

How Did This Get Funded?!
Automatically Identifying Quirky Scientific Achievements

Core Question:

Can we automatically detect funny and unusual scientific papers?

Shani et al. 2021

Dataset

Scientific Paper Titles

- 1707 humorous papers
  - 211 Ig Nobel winners
  - Others manually collected from online forums and blogs
- 1707 randomly sampled papers
  - Fields: neuroscience, medicine, biology, exact sciences
  - Select to preserve field balance

Binary Labels

Is this paper title humorous / "Ig Nobel worthy"?

Shani et al. 2021

Classification

Feature Categories:

- Unexpected Language
- Simple Language
- Crude Language
- Funny Language

Model	Accuracy	Precision	Recall
Iggy	0.897	0.901	0.893
SciBERT	0.910	0.911	0.911
SciBERT^f	**0.922**	**0.919**	**0.926**
BERT	0.904	0.906	0.893
BERT^f	0.900	0.899	0.902
RF	0.761	0.746	0.796
LR	0.781	0.754	0.837

Table 2: Accuracy of the different models on our dataset using cross validation with k=5. SciBERT^f outperforms.

Shani et al. 2021

Evaluating “in the Wild”

Title	Models
The kinematics of eating with a spoon: Bringing the food to the mouth, or the mouth to the food?	Iggy, BERT^f, SciBERT^f
Do bonobos say NO by shaking their head?	Iggy, BERT^f, SciBERT^f
Is Anakin Skywalker suffering from borderline personality disorder?	Iggy, BERT^f, SciBERT^f
Not eating like a pig: European wild boar wash their food	Iggy, BERT^f
Why don’t chimpanzees in Gabon crack nuts?	SciBERT^f, BERT^f
Why do people lie online? “Because everyone lies on the internet”	BERT^f
Which type of alcohol is easier on the gut?	BERT^f
Rainbow connection and forbidden subgraphs	BERT
A scandal of invisibility: making everyone count by counting everyone	SciBERT
Where do we look when we walk on stairs? Gaze behaviour on stairs, transitions, and handrails	SciBERT

Table 4: A sample of top rated papers found by our models.

Shani et al. 2021

Unravelling Names of Fictional Characters

Core Question:

Can the polarity of a character's role be predicted by their name alone?

"Hero"

Iago? "Villain"

Papantoniou et al. 2016

Kevinism

"Kevin isn't a name, but a diagnosis"

"Kevin ist kein Name, sondern eine Diagnose"

Trenkamp 2009

Dataset

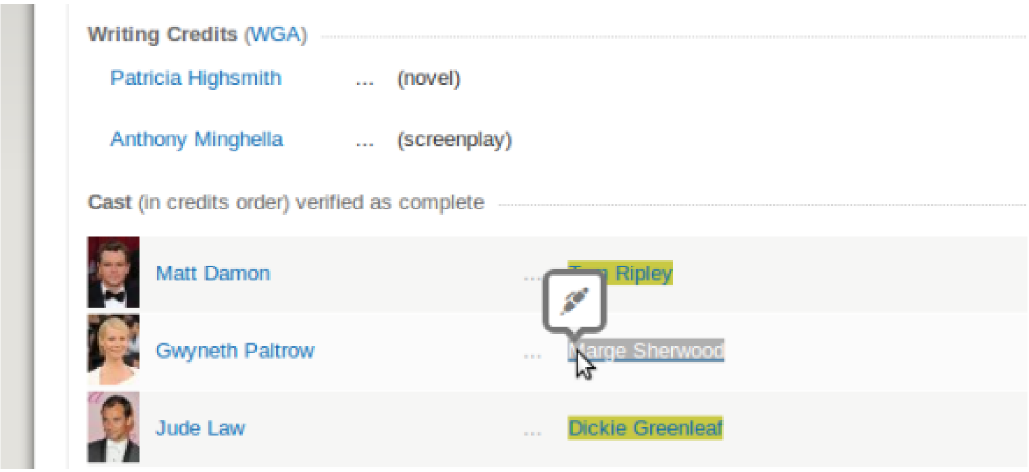


Figure 1: Character annotation tool

- Positive: the role of the character in the plot left a positive impression
- Negative: the role of a character left a negative impression
- *Neutral: ignored label*

Papantoniou et al. 2016

Results

" feature " ablation

	Rec.	Prec.	F-score
Without domain features	0.803	0.801	0.802
Only domain features	0.725	0.699	0.667
Only phonological features	0.790	0.786	0.787
Without poetic features	0.836	0.832	0.833
Without consonance feature	0.823	0.820	0.821
Without emotions features	0.814	0.810	0.811
Without phonological features	0.798	0.792	0.793
Without social features	0.807	0.803	0.804
All features	0.824	0.822	0.823

Table 5: Performance of J48 for different feature settings

Papantoniou et al. 2016

Feature Analysis

Phonemes	Class
/p/, /b/ (bilabial plosive)	P
/l/ (alveolar lateral)	P
/f/, /v/ (labiodental africative)	N
/k/, /g/ (velar plosive)	N
/t/, /d/ (alveolar plosive)	N
/dʒ/, /tʃ/ (affricate)	N
/m/, /n/ (nasal)	N
/ɹ/ (alveolar retroflex)	N

Table 7: Consonants behavior

Most frequent in positive characters	
Phoneme n-gram	Examples
/lɪ/	Ned Alleyn (Shakespeare in Love)
/an/	Anouk Rocher (Chocolat)
/aɪ/	Eliza Doolittle (My Fair Lady)
/nɪ/	Linguini (Ratatouille)
/ɪst/	Kevin McCallister (Home Alone)
/ɹəʊ/	Frodo (The Lord of the Rings)
/and/	Dylan Sanders (Charlie's Angels)
/stə/	C.C. Baxter (The Apartment)
Most frequent in negative characters	
Phoneme n-gram	Examples
/ən/	Tom Buchanan (The Great Gatsby)
/əʊ/	Iago (Aladdin)
/tə/	Norrington (Pirates of the Caribbean)
/ɹɪ/	Tom Ripley (The Talented Mr. Ripley)
/mən/	Norman Bates (Psycho)
/mɪs/	Mystique (X-Men)
/ktə/	Hannibal Lecter (Hannibal)

Table 6: Frequent phoneme {2,3}-grams

Papantoniou et al. 2016