Supplemental Material for "It's going to be okay: Measuring Access to Support in Online Communities"

Zijian Wang
Symbolic Systems Program
Stanford University
zijwang@stanford.edu

David Jurgens
School of Information
University of Michigan
jurgens@umich.edu

Abstract

This document contains supplemental materials describing the classifiers and additional results. This paper has an accompanying data and code release as well.

1 Support Classifier

Here, we describe the features used in the classifier and provide additional supplemental analyses.

1.1 Classifier Features

Prior to feature extraction, we minimally normal text by standardizing whitespace to one space at most. We also replace common misspellings with their normalized variants. We report all of these in the associated code release.

Sentence Features

- Binary feature of whether the sentence is all lower case
- Binary features of whether the sentence starts with a capitalized letter

Parse Features We use spaCy to parse each sentence in a post into (governor, relation, dependent) tuples. We include these as 1-hot features, keeping only those dependency features that occur at least 5 times in the data.

- Each triple, where words are replaced by their parts of speech
- Each two-element combination of the triple, where words are replaced by part of speech.

Stylistic Features

- Number of capitalized words, excluding "I"
- Number of all-CAPS words
- Length of longest all-CAPS span
- Percent of verbs in passive constructions

Content Features

- Number of hedges
- Number of first-person pronouns
- Length of third-person pronouns
- Percent of verbs in passive constructions
- Number of curse words
- Number of laughs
- Number of disfluencies

Distributional Features

- Average word2vec vector for all words (Mikolov et al., 2013), using the public Google News vectors.
- Average word2vec vectors nouns, verbs, adjectives, and adverbs, each computed separately using the public Google News vectors.

Content Analysis Features

- Percent of sentences where the sentence is positive, calculated using TextBlob (Loria et al., 2014)
- Average subjectivity rating of the sentences, calculated using TextBlob (Loria et al., 2014)
- Average word formality, rated using the data of Pavlick and Nenkova (2015)
- Average word frequency, computed from Google N-Grams, excluding stopwords
- Average Flesch-Kincaid Grade Level across sentences

Syntactic Features Part of speech tags are calculated by spaCy.

- Normalized part of speech frequency

Lexicon Features We release all lexicons used in the paper in the associated code release, with the exception of LIWC. However, we note that in practice Empath provides similar performance and may obviate the need for LIWC as a separate lexicon.

- LIWC (Pennebaker et al., 2001)
- Liu et al. (2005)
- Empath (Fast et al., 2016)
- Argumentation Phrases (Teufel, 2000).
- Word concreteness ratings (Brysbaert et al., 2014)
- NRC emotion lexicons (Mohammad and Turney, 2013)

Length Features

- Mean word length in characters
- Mean sentence length in words
- Mean sentence length in characters

Other Features

- All features reported in Danescu-Niculescu-Mizil et al. (2013)
- All features reported in Pavlick and Tetreault (2016), except for those involving named entities.

1.2 Analysis

Cross-platform performance for the Support classifier is shown in Table 1 and reveals that StackExchange is largely responsible for lower performance overall. However, Wikipedia and Reddit-trained models perform equally poor on data from each other's platform. Nevertheless, all cross-platform models still substantially outperformed the random chance and most-frequent label baselines on any one platform.

		Test Data		
		Reddit	StackExchange	Wikipedia
Train Data	Reddit	0.5366	0.4037	0.3958
	SE	0.3710	0.4433	0.3718
	Wiki.	0.3942	0.4202	0.5340
	Random	0.2678	0.2502	0.2661
	Most Freq.	0.2854	0.2978	0.2939

Table 1: Cross-platform performance of the support classifier

2 Support Annotations

Table 2 shows additional longer examples of supportiveness annotations from our dataset.

3 Gender Classifiers

3.1 Gender in Names

Table 3 and 4 show the result of comparison experiments using the dataset and evaluation metrics from Knowles et al. (2016) and Jaech and Ostendorf (2015). Note that we do not include the data they used to train their models in our own training data. Thus, while the data is in-domain for their evaluations is predicting on out of domain names. Despite this difference, our model outperforms both consistently.

Figure 5 shows the relationship between Macro-F1 precision and recall (the percentage of data left after applying the threshold) when adjusting the threshold in the gender identification model. Note that our the *high confidence* predictions used in the paper occur at the 0.4 threshold, which has a precision of more than 0.9.

All data used to classify gender performance in names was partitioned into 80% train, 10% development, and 10% test splits. For Twitter, each user contributes both a screen name and a username. In the setup described in §4.1, these Twitter names were partitioned randomly across the splits. However, an individual could style their username and screen name in a similar fashion, which raises the potential for data leakage between training and test if a user's two names are in different splits. While this potential leakage does not affect the validity of the classifier, it could overestimate the performance of the classifier. Therefore, as a follow-up test, we repeat the setup of §4.1 and first partition users into one of the test, train, or development splits, then assigning both of their names to the split. To compare both models, we test on the subset of users from the first model where both names were in the test data (52K users total). The resulting performance is 0.8016 F1 for the model using the setup described in §4.1 and 0.8099 for the model using the revised setup. This difference is significant at $p<0.01$ using McNemar's test of significance indicating there is sufficient regularity between screen name and username that the performance in Table 4 for Twitter is an overestimate (performance for Reddit is unchanged); however, the small magnitude of the difference between the models suggests the overestimate is relatively close to the true performance.

[1]Based on 57,142 still public available unique users out of 58,046 unique users mentioned in Knowles et al. (2016).

[2]We are using the F1 defined in Knowles et al. (2016) as "the harmonic mean of accuracy and coverage" for consistency.

Rating	Example
1.33	see your arse mate, stop talking out of it.
1.33	thats total bullshit. its possible that thats what hes doing, but you dont know anything. especially not about kyrie who seems to be the most unpredictable dude in the nba
1.66	Everything you said here is complete nonsense. Seriously. You sound like a jackass.
1.66	If your adult child despises you for implementing house rules and expecting them to abide by them *when theyre there as a fucking guest* then you raised some bratty ass kid
2.0	Following up, you adding more sources to something that isn't even debated (the claim) is just WP:OVERCITE.
2.0	Have you even read the series?
2.33	I wont answer your idiotic question, but I have a question for you. If you dont see anything wrong with killing 300 people and then stealing their bodies, dont you think it s time to rethink your attitude?
2.33	Apparently you dont understand the concept of punching up vs punching down.
2.66	I have no idea what youre going for.
2.66	Guess well just have to turn off the sun, too, then, wont we?
3.0	I agree with the above statement.
3.0	I don't know if you were referring to this or not, but in case you hadn't heard of it ... <link>
3.0	depends on whether they know that's what she's doing.
3.33	Damn, this makes me feel better :)
3.66	Kindly edit the title. Your suggestion is correct
3.66	Provided one expands it and one justifies every step, indeed this is a good basis for an answer.
4.0	That means a lot coming from you. Thank you for the help, as well.
4.0	You should write this as an answer.
4.33	most surely not. I think its a great question and I have been asking myself the same question while looking at some old documents on the web.
4.33	Awesome! Havent gotten the chance to go sky diving but Ive been to Vegas and trust me when I say youll be in for a treat

Table 2: Additional examples of annotator ratings of Support from 1 to 5 (3.0 is neutral).

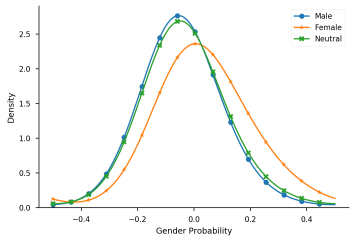

Figure 1: Reddit

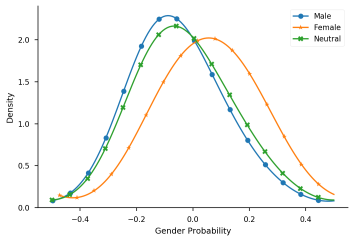

Figure 2: StackExchange

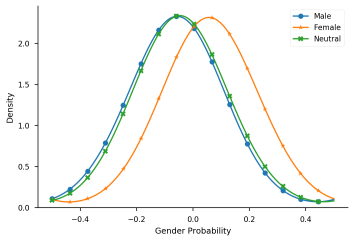

Figure 3: Wiki

Figure 4: The distribution of gender probabilities as predicted from the text of a post, centered to $[-0.5, 0.5]$ where 0 denotes no discernible gender. Separate curves are shown relative to the inferred gender label of the user on the basis of their name (from held out data not used for training), with subfigures for each platform.

3.2 Gender from Text

Features The feature set for the text-based classifier is largely a subset of the feature set used for support with the exceptions:

- No parsing features
- No part of speech features
- No politeness features
- No support features

Models	Demographer Data				Our Test Data					
	Wiki		Twitter[1]		Reddit		Twitter		Merged	
	Coverage	F1[2]	Coverage	F1	Coverage	F1	Coverage	F1	Coverage	F1
Our Model	**1.0**	**0.9626**	**1.0**	**0.9406**	**1.0**	**0.7834**	**1.0**	**0.8797**	**1.0**	**0.8762**
Demographer	0.9999	0.9497	0.9874	0.9052	0.9952	0.7178	0.9980	0.7894	0.9978	0.7867

Table 3: Comparison with Demographer (Knowles et al., 2016)

Models	Jaech and Ostendorf (2015)'s Data	Our Test Data		
	OkCupid	Reddit	Twitter	Merged
	Accuracy			
Our Model	**0.7671**	**0.6439**	**0.7852**	**0.7796**
Jaech and Ostendorf (2015)	0.7421	0.6037	0.7101	0.7059

Table 4: Comparison with Jaech and Ostendorf (2015)

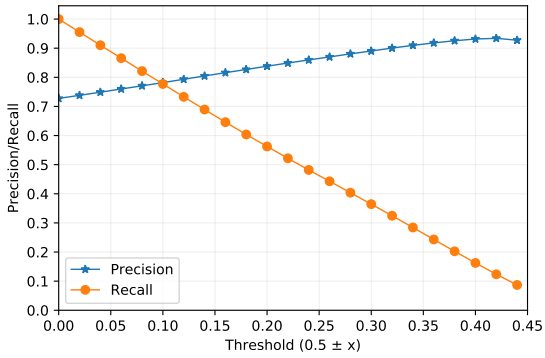

Figure 5: Precision-Recall Curve of the Gender Identification Model

The gender from text classifier adds one new feature for the average number of numbers used in text.

Due to the high dimensional space for the regression, we limit the total number of n-gram features to the most frequent 10,000 1-gram, 2-gram, and 3-gram seen for each. All other features must occur at least 5 times to be included.

As an example of the difficulty of the task, we include example posts from each platform in Table 5 and the inferred gender from the username.

Additional Results Figure 4 shows the distribution of gender probabilities as predicted from the text of a post. Note that the distributions for all three gender labels are closely centered near zero, highlighting the fact that most posts by any user do not strongly convey gender. However, male- and female-named users do have their posts shifted slightly towards the direction of their respective gender indicating that users who choose names associated with one gender have a *slight* tendency to write in a way that conveys that gender. We also note that the neutral gender category is more similar to users with male names; as all three platforms have higher male populations, it should be expected that a random sample of neutral names would have more (neutrally-named) men, which likely accounts for the increased similarity the writing style for with male-named users.

Source	Post	Gender
Reddit	I'm keeping the radishes	Female
Reddit	All right, well, you don't have an example of it being found unconstitutional, and we're otherwise just going in circles. Suffice it to say, we disagree. So be it.	Female
Reddit	Thank you to anyone who has had their best Type-0 unit as their friend unit. And another thank you to anyone who keeps their best Type-0 unit as their friend unit for the remainder of the event. You are appreciated.	Male
SE	Well as I understand it the primality test for prime numbers does a modulo function with the Messene prime being tested but either way I'd like to know the largest number that can be computed and handled using our current level of technology (I know it is continually improving)	Female
SE	While there is no clear evidence that Buckbeak was killed, it seems like there was a clear view of the execution seeing as Draco & Co were at that vantage point to watch. As well, while our POV doesn't actually view the execution, if Harry, Ron and Hermione were watching, they'd clearly see that Macnair chopped the pumpkin as opposed to a giant hipogryph, and their facial expressions and moment of crying/sadness afterwards would suggest that either Buckbeak was killed, or they REALLY loved that pumpkin.	Male
SE	Thank you for your answer :) can I write the values (1.454, 2.14,4.23) on the graph ?	Female
Wiki	This article could be improved if someone knowledgeable of the subject would insert a section with examples, preferably of all types of generalized inverse addressed in the article.	Female
Wiki	There are two references in the "Origin of trials" section that refer to "Wiccan sources" which feel rather out of place in this largely historical article. I know that not every reference must be attributed, but without any more specific description of "Wiccan sources" these two references seem questionable at best, especially in a good article nominee.	Male
Wiki	In the Shiloh section it says Grant's General Order No. 11, barring Jews from cotton trading, was "anti-semitic". But was this just a quick fix, in response to the high numbers of Jewish merchants "profiteering from an illicit cotton exchange through enemy lines while Union soldiers died in the fields"? Or did Grant actually hate Jews? This is sort of a controversial item, and the only opinion we have on that note is from Jean Edward Smith, who as I said, is mentioned by name three times in the text. Is there any other significant view to balance out Smith's opinion, or should we just get rid of Smith's personal viewpoint altogether? Unless we can show hatred and can refute Grant's concern for illicit trading via enemy lines, I'd recommend getting rid of this example of 20th century hyper-speak and let the readers decide this sort of thing for themselves. Statements like this undermine the meaning of real anti-semitism.	Male

Table 5: Examples of post from individuals with high-confidence gender predictions.

References

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46(3):904–911.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4647–4657. ACM.

Aaron Jaech and Mari Ostendorf. 2015. What your username says about you. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2032–2037.

Rebecca Knowles, Josh Carroll, and Mark Dredze. 2016. Demographer: Extremely simple name demographics. In *EMNLP Workshop on NLP and Computational Social Science*, pages 108–113. Association for Computational Linguistics.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 342–351. ACM.

Steven Loria, P Keen, M Honnibal, R Yankovsky, D Karesh, E Dempsey, et al. 2014. Textblob: simplified text processing. *Secondary TextBlob: Simplified Text Processing*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, pages 3111–3119.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Artificial Intelligence*, 29(3):436–465.

Ellie Pavlick and Ani Nenkova. 2015. Inducing lexical style properties for paraphrase and genre differentiation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 218–224.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association of Computational Linguistics (TACL)*, 4(1):61–74.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Simone Teufel. 2000. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, University of Ediburgh.