A Dataset and Classifier for Recognizing Social Media English

Su Lin Blodgett* **Johnny Tian-Zheng Wei†** **Brendan O'Connor***
University of Massachusetts Amherst, Amherst, MA
*{blodgett, brenocon@cs.umass.edu} †jwei@umass.edu

Abstract

While language identification works well on standard texts, it performs much worse on social media language, in particular dialectal language—even for English. First, to support work on English language identification, we contribute a new dataset of tweets annotated for English versus non-English, with attention to ambiguity, code-switching, and automatic generation issues. It is randomly sampled from all public messages, avoiding biases towards pre-existing language classifiers. Second, we find that a demographic language model—which identifies messages with language similar to that used by several U.S. ethnic populations on Twitter—can be used to improve English language identification performance when combined with a traditional supervised language identifier. It increases recall with almost no loss of precision, including, surprisingly, for English messages written by non-U.S. authors.

Our dataset and identifier ensemble are available online.[1]

1 Introduction and Related Work

Language identification is the task of determining the major world language a document is written in. A range of supervised classification methods—often based on character n-gram features—achieve excellent performance for this problem on long, monolingual documents (Hughes et al., 2006). But short documents are much more challenging, such as Twitter messages (Lui and Baldwin, 2012, 2014; Bergsma et al., 2012; Williams and Dagli, 2017).

Compounding the challenge is domain mismatch: the types of casual language, dialectal language, and Internet-specific constructs found in social media are often not present in the standardized genres of training data for existing language identifiers. This is potentially especially problematic for language by minority dialect speakers—for example, Blodgett et al. (2016) found that current language identification models had lower recall for tweets written in African-American English (AAE) than those in standard English. This is not surprising given the domain mismatch—a survey of recent language identifiers shows that common sources of training data are Wikipedia, newswire (e.g. the Leipzig corpora), and government and legal documents such as EuroGov, EuroParl, or the Universal Declaration of Human Rights (Lui and Baldwin, 2012; King and Abney, 2013; Jaech et al., 2016; Kocmi and Bojar, 2017; Lui and Cook, 2013).

A language identification system typically aims to classify messages as one of a few hundred major world languages, which are generally well-resourced mainstream language varieties with officially recognized status by major political entities; these language varieties typically have official ISO 639 codes assigned to them (which are returned by language identification software APIs).[2] Given the high linguistic diversity of messages in social media, it is tempting to imagine fine-grained dialect identification (for example, identifying messages written in AAE), but at the same time, the traditional task of identifying major world languages will continue to be useful (for example, an AAE message could be reasonably analyzed with general English language technologies). In this work we maintain the paradigm of treating English as a broad language category, but propose that the texts

[1] http://slanglab.cs.umass.edu/TwitterLangID

[2] For example, langid.py, CLD2, Microsoft Azure, IBM Watson, and Google Translation API all offer ISO-returning language identification software or services.

Proceedings of the 3rd Workshop on Noisy User-generated Text, pages 56–61
Copenhagen, Denmark, September 7, 2017.

that match it ought to be broadened to include non-standard, social media, and dialectal varieties of English.

If there was abundant language-annotated Twitter data, it would be straightforward to train an in-domain language identifier. But very little exists, since it is inherently time-consuming and expensive to annotate. Datasets are typically small, or semi-automatically tagged (Bergsma et al., 2012), which may bias them towards pre-existing standardized language.

A promising approach is to leverage large quantities of non-language-labeled tweets to help adapt a standard identifier to perform better on social media. If the messages are treated as unlabeled, this could be framed as unsupervised domain adaptation problem, for which a number of approaches are available (Blitzer et al., 2006, 2007; Plank, 2009; Yang and Eisenstein, 2016).

We focus on a unique, and different, large-scale training signal—U.S. neighborhood-level demographics. There is considerable linguistic diversity within the U.S., and its geographic patterns have some rough correlation with different ethnic and race populations. Blodgett et al. analyzed them with a mixed membership model—for which messages written by authors living in areas heavy in a particular demographic group were more likely to use a unigram language model associated with that group—in order to focus on AAE. But they note their model found that non-English language tended to gravitate towards one of the latent language models, which was useful to better identify English spoken within the U.S. that a standard identifier missed.

We hypothesize that this generalizes beyond specific dialect populations within the U.S., testing whether this soft signal from the demographic model actually gives a better model of overall social media English. We evaluate as fairly and completely as possible; we first annotate a new dataset of uniformly sampled tweets for whether they are English versus non-English (§2). In §3, we apply Blodgett et al.'s model to infer U.S. demographic language proportions in new tweets, finding that when added as an ensemble to a pre-existing identifier, performance improves—including when paired with feature-based, neural network, and proprietary identifiers. Such ensembles perform better than in-domain training with the largest available annotated Twitter dataset, and also better than a self-training domain adaptation approach on the same dataset used to construct the demographic language model—and the accuracy increases for English messages from many different countries around the world.

Label	Full Count	Evaluation Count
English	5086	3758
Not English	4646	4608
Ambiguous	770	0
Total	10502	8366

Table 1: Dataset statistics for each language label; the evaluation count refers to the subset used for evaluation.

Label	Count
Code-Switched	162
Ambiguous due to Named Entities	132
Automatically Generated	1371

Table 2: Dataset statistics for additional labels.

2 Dataset and Annotation

We sampled 10,502 messages from January 1, 2013 to September 11, 2016 from an archive of publicly available geotagged tweets. We annotated the tweets with three mutually exclusive binary labels: *English*, *Not English*, and *Ambiguous*. These tweets were further annotated with descriptive labels:

- *Code-switched*: Tweets containing both text in English and text in another language.
- *Ambiguous due to named entities*: Tweets containing only named entities, such as *Vegas!*, and therefore whose language could not be unambiguously determined.
- *Automatically generated*: Tweets whose content appeared to be automatically generated, such as *I just finished running 15.21 km in 1h:17m:32s with #Endomondo #endorphins https://t.co/bugbJOvJ31*.

We excluded any usernames and URLs in a tweet from the judgment of the tweet's language, but included hashtags. Tables 1 and 2 contain the statistics for these labels in our annotated dataset. For all our experiments, we evaluate only on the subset of messages in the dataset not labeled as ambiguous or automatically generated, which we call the evaluation dataset.

3 Experiments

3.1 Training Datasets

We investigate the effect of in-domain and extra out-of-domain training data with two datasets. The first is a dataset released by Twitter of 120,575 tweets uniformly sampled from all Twitter data, which were first labeled by three different classifiers (Twitter's internal algorithm, Google's Compact Language Detector 2, and *langid.py*), then annotated by humans where classifiers disagreed.[3] We reserve our own dataset for evaluation, but use this dataset for in-domain training. This dataset is only made available by tweet ID, and many of its messages are now missing; we were able to retrieve 74,259 tweets (61.6%). For the rest of this work, we call this the Twitter70 dataset (since it originally covered about 70 languages).

In addition, following Jaech et al. (2016), we supplemented Twitter70 with out-of-domain Wikipedia data for 41 languages,[4] sampling 10,000 sentences from each language.

3.2 Classifiers

We tested a number of classifiers on our annotated dataset trained on a variety of domains, and in some cases retrained.

- CLD2: a Naive Bayes classifier with a pretrained model from a proprietary corpus; it offers no support for re-training.
- Twitter: the output of Twitter's proprietary language identification algorithm.
- *langid.py*: a Naive Bayes classifier for 97 languages with character n-gram features, including a pretrained model based on text from JRC-Acquis, ClueWeb 09, Wikipedia, Reuters, and Debian i18n (Lui and Baldwin, 2012).
- Neural model: a hierarchical neural classifier that learns both character and word representations. It provides a training dataset with 41,250 Wikipedia sentence fragments in 33 languages (Jaech et al., 2016).[5]

Self-training We experimented with one simple approach to unsupervised domain adaptation: self-training with an unlabeled target domain corpus (Plank, 2009) by using *langid.py* to label the corpus of tweets–released by Blodgett et al.[6] and the same one used to train their demographic model–then collecting those tweets classified with posterior probability greater than or equal to 0.98. We downsampled tweets classified as English to 1 million, yielding a total corpus of 2.2 million tweets. Since we did not have access to *langid.py*'s original training data, we trained a new model on this data, then combined it as an ensemble with the original model, where a tweet was classified as English if either component classified it as English.

Demographic prediction ensemble Blodgett et al. describes applying a U.S. demographically-aligned language model as an ensemble classifier, using a mixed membership model trained over four demographic topics (African-American, Hispanic, Asian, and white). For this classifier, tweets are first classified by an off-the-shelf classifier; if it is classified as English, the classification is accepted. Otherwise, the off-the-shelf classifier is overriden and the tweet classified as English if the total posterior probability of the African-American, Hispanic, and white topics under the demographic model was at least 90%. Table 3 lists these ensembles as "+ Demo". Blodgett et al. found the classifier seemed to improve recall, but this work better evaluates the approach with the new annotations.

3.3 Length-Normalized Analysis

From manual inspection, we observed that longer tweets are significantly more likely to be correctly classified; we investigate this length effect by grouping messages into five bins (shown in Table 6) according to the number of words in the message. We pre-processed messages by fixing HTML escape characters and removing URLs, @-mentions, emojis, and the "RT" token. For each bin, we calculate recall of the *langid.py* and the demographic ensemble classifier with *langid.py*.

4 Results and Discussion

We evaluated on the 8,366 tweets in our dataset that were not annotated as ambiguous or automatically generated. Table 3 shows the precision and recall for each experiment. We focus on recall, as Blodgett et al.'s analysis indicates that while precision is largely consistent across experiments, there is a significant gap in recall performance across different varieties of English.

[3] https://blog.twitter.com/2015/evaluating-language-identification-performance

[4] https://sites.google.com/site/rmyeid/projects/polyglot

[5] Kocmi and Bojar (2017) offer an alternative neural model for language identification.

[6] http://slanglab.cs.umass.edu/TwitterAAE

Model	Training	Precision	Recall
CLD2	(1) Pre-trained	0.948	0.863
	(2) + Demo.	0.946	0.924 (+ 6.1%)
Tw. internal	(3) Pre-trained	0.979	0.866
	(4) + Demo.	0.974	0.925 (+ 5.9%)
langid.py	(5) Pre-trained	0.923	0.886
	(6) + Vocab.	0.472	0.993
	(7) Self-trained	0.924	0.894
	(8) + Demo.	0.923	0.930 (+ 3.6%)
	(9) Twitter70	0.927	0.940
	(10) + Demo.	0.923	0.957 (+ 1.7%)
	(11) Tw70 and Wiki.	0.946	0.903
	(12) + Demo.	0.943	0.946 (+ 4.3%)
Neural	(13) Pre-trained	0.973	0.415
	(14) + Demo.	0.976	0.773 (+ 35.8%)
	(15) Twitter70	0.949	0.840
	(16) + Demo.	0.946	0.892 (+ 5.2%)

Table 3: English classification results on not ambiguous, not automatically generated tweets. "+ Demo." indicates including in an ensemble with the demographics-based English classifier.

Country	En	∼En	*langid.py* Recall	Ens. Recall
USA	2368	80	0.968	0.982
Brazil	42	945	0.833	0.833
Indonesia	161	707	0.764	0.767
Turkey	13	304	0.769	0.846
Japan	14	340	0.929	1.0
United Kingdom	401	18	0.962	0.980
Malaysia	90	174	0.833	0.833
Spain	28	263	0.75	0.821
Argentina	10	291	0.7	0.7
France	26	206	0.846	0.846
Mexico	25	162	0.76	0.76
Philippines	91	86	0.934	0.945
Thailand	14	111	0.643	0.786
Russia	9	129	0.667	0.778
Canada	96	7	0.979	0.990

Table 4: Language counts for countries with at least 100 non-ambiguous, non-automatically generated messages (out of 129 countries total), with English recall for the best-performing *langid.py* model and that model in an ensemble classifier.

Tweet
*@username good afternoon and Happy Birthdayyyyyyyyyy *Turns on music* Time to partyyyyy*
I miss you! #vivasantotomas #ust #goUST #igers #igdaily #igersasia #igersmanila #instagood
Sooo fucked yuuuuppp bouuutta start a figgght
catch mines you catch yours we both happy...
Go follow me on Instagram @username and like 5 pics for a goodmorningg post
Think me & my baddies getting rooms dis weekend!
@username HML if u do B
@username @username FR LIKE I CANT EVEN DEAL WITH PEOPLE LIKE THIS
I k you dont like me lowkey but hey
@username I DORN WVEN WTCH GIRL MEETS WORLDBUT IM WATCHINF THAT EPISODE

Table 5: Sample of tweets which were mis-classified as non-English by *langid.py* but correctly classified by the demographic ensemble. @-mentions are shown as @username for display in the table.

Unsurprisingly, we found that training on Twitter data improved classifiers' English recall, compared to their pre-trained models. In our experiments, we found that recall was best when training on the subset of the Twitter70 dataset containing only languages with at least 1,000 tweets present in the dataset. We also found that the additional information provided by the demographic model's predictions still adds to the increased performance from training on Twitter data. Notably, precision decreased by no more than 0.4% when the demographic model is added.

We also noted that pre-processing improved recall by 1 to 5%.

Proprietary algorithms We found that neither CLD2 nor Twitter's internal algorithm was competitive with *langid.py* out of the box, in line with previous findings, but combining their predictions with demographic predictions did increase recall.[7]

langid.py Self-training *langid.py* produced little change compared to the original pre-trained model (rows (5) vs. (7)), despite its use of 2.2 million new tweets from self-training step. We observed that even tweets that *langid.py* classified as non-English with more than 0.98 posterior probability were, in fact, generally English. This suggests that tweets are sufficiently different from standard training data that it is difficult for self-training to be effective. In contrast, simple in-domain training was effective: retraining it with the Twitter70 dataset achieved substantially better recall with a

[7] We tried several times to run the Google Translate API's language identifier, but it returned an internal server error for approximately 75% of the tweets.

5.4% raw increase compared to its out-of-domain original pretrained model (rows (5) vs. (9)).

In all cases, regardless of the data used to train the model, *langid.py*'s recall was improved with the addition of demographic predictions; for example, the demographic predictions added to the pre-trained model brought recall close to the model trained on Twitter70 alone, indicating that in the absence of in-domain training data, the demographic model's predictions can make a model competitive with a model that does have in-domain training data (rows (8) vs. (9)). Of course, in-domain labeled data only helps more (10).

Neural model Finally, the neural model performed worse than *langid.py* when trained on the same Twitter70 dataset (rows (9) vs. (15)), and its performance lagged when trained on its provided dataset of Wikipedia sentence fragments.[8] As with the other models, demographic predictions again improve performance.

Table 5 shows a sample of ten tweets misclassified as non-English by *langid.py* and correctly classified by the demographic ensemble as English. Several sources of potential error are evident; many non-conventional spellings, such as *partyyyyy* and *watchinf*, do not challenge an English reader but might reasonably challenge character n-gram models. Similarly, common social abbreviations such as *hml* and *fr* are challenging.

4.1 Improving English Recall Worldwide

We further analyzed our English recall results according to messages' country of origin, limiting our analysis to countries with at least 100 non-ambiguous, non-automatically generated messages in our dataset. For each country's messages, we compared the recall from best standalone *langid.py* model (trained on Twitter70) and the recall from same model combined with demographic predictions, as shown in Table 4. Surprisingly, for ten of the fifteen countries we found that using demographic predictions improved recall performance, suggesting that the additional soft signal of "Englishness" provided by the demographic model aids performance across tweets labeled as English globally. In future work, we would like to investigate linguistic properties of these non-U.S. English tweets.

[8] Unfortunately, we were unable to train it on the same Wikipedia data as in (11), which is a bit larger.

	Message Length	*langid.py* Recall	Ensemble Recall
English	$t \leq 5$	80.7	91.9
	$5 < t \leq 10$	88.8	92.4
	$10 < t \leq 15$	91.9	93.0
	$15 < t \leq 20$	96.1	96.7
	$t \geq 20$	97.2	97.5
Non-English	$t \leq 5$	90.0	99.9
	$5 < t \leq 10$	95.2	99.5
	$10 < t \leq 15$	95.6	99.9
	$15 < t \leq 20$	95.2	1.0
	$t \geq 20$	95.2	1.0

Table 6: Percent of the messages in each bin classified correctly as English or non-English by each classifier; t is the message length for the bin.

4.2 Improving Recall for Short Tweets

Our results from the length-normalized analysis, shown in Table 6, demonstrate that recall on short tweets, particularly short English tweets, is challenging; unsurprisingly, recall increases as tweet length increases. More importantly, for short tweets the demographic ensemble classifier greatly reduces this gap; while the difference in *langid.py*'s recall performance between the shortest and longest English tweets is 16.5%, the difference is only 5.6% for the ensemble classifier. The gap is similarly decreased for non-English tweets. We note also that precision is consistently high across all bins for both *langid.py* and the ensemble classifier. The experiment indicates that the demographic model's signal of "Englishness" may aid performance not only for global varieties of English, but also for short messages of any kind.

5 Conclusion

In this work, we presented a fully human-annotated dataset and evaluated a range of language identification models in a series of experiments across training datasets and in-domain and domain adaptation settings. We find that predictions from a partially supervised demographic model aids in recall performance across tweets labeled as English drawn from a range of countries, particularly in the absence of in-domain labeled data; we hope that our dataset will aid research in international varieties of English (Trudgill and Hannah, 2008). In future work, we would like to investigate other domain adaptation approaches; in addition, we would like to adapt the demographic model to other languages where dialectal

variation might present similar challenges.

References

Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clayton Fink, and Theresa Wilson. 2012. Language identification for creating language-specific twitter collections. In *Proceedings of the second workshop on language in social media*, pages 65–74. ACL.

John Blitzer, Mark Dredze, Fernando Pereira, et al. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, volume 7, pages 440–447.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of EMNLP*. ACL.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of EMNLP*, Austin, Texas. ACL.

Baden Hughes, Timothy Baldwin, Steven Bird, Jeremy Nicholson, and Andrew MacKinlay. 2006. Reconsidering language identification for written language resources. In *Proceedings of LREC*. European Language Resources Association.

Aaron Jaech, George Mulcaire, Shobhit Hathi, Mari Ostendorf, and Noah A. Smith. 2016. Hierarchical character-word models for language identification. In *Proceedings of EMNLP*, Austin, TX, USA. ACL.

Ben King and Steven P Abney. 2013. Labeling the languages of words in mixed-language documents using weakly supervised methods. In *Proceedings of NAACL-HLT*.

Tom Kocmi and Ondřej Bojar. 2017. Lanidenn: Multilingual language identification on character window. *To appear in Proceedings of EACL*.

Marco Lui and T. Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012), Demo Session, Jeju, Republic of Korea*.

Marco Lui and Timothy Baldwin. 2014. Accurate language identification of twitter messages. In *Proceedings of the 5th workshop on language analysis for social media (LASM)@ EACL*, pages 17–25.

Marco Lui and Paul Cook. 2013. Classifying english documents by national dialect. In *Proceedings of the Australasian Language Technology Association Workshop (ALTA)*.

Barbara Plank. 2009. A comparison of structural correspondence learning and self-training for discriminative parse selection. In *Proceedings of the NAACL HLT Workshop on Semi-supervised Learning for NLP*. ACL.

Peter Trudgill and Jean Hannah. 2008. *International English: A guide to varieties of Standard English*. Routledge.

Jennifer Williams and Charlie Dagli. 2017. Twitter language identification of similar languages and dialects without ground truth. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. ACL.

Yi Yang and Jacob Eisenstein. 2016. Part-of-speech tagging for historical English. In *Proceedings of NAACL-HLT*, San Diego, California. ACL.