Efficacy of Named Entity Recognition in Text Messages

By Leland Ball

December, 2020

Table of Contents

Introduction

Named entity recognition (NER) is the ability to identify sets of words in unstructured text as belonging to various categories. These categories include persons, organizations, quantities, locations and so on. One of the major focuses of NER is in the context of free-text found on the web. This can be seen by looking at what data certain NER-capable Natural Language Processing (NLP) models are trained on, which is often scrapings from websites or reviews posted to the web. While not necessarily overly formal in structure, this common training data is often more formal than another body of unstructured data: text messages. In this paper the efficacy of these NER algorithms and models are analyzed for their performance upon the highly contextualized and informal speech found in cellphone text messages. In this analysis, a hand-coded set of text messages are used to compare the NER algorithm provided by the Spacey NLP library, and a naïve baseline list of names. Different accuracy metrics are compared, including accuracy, precision, recall, and F1 scores.

Research Question

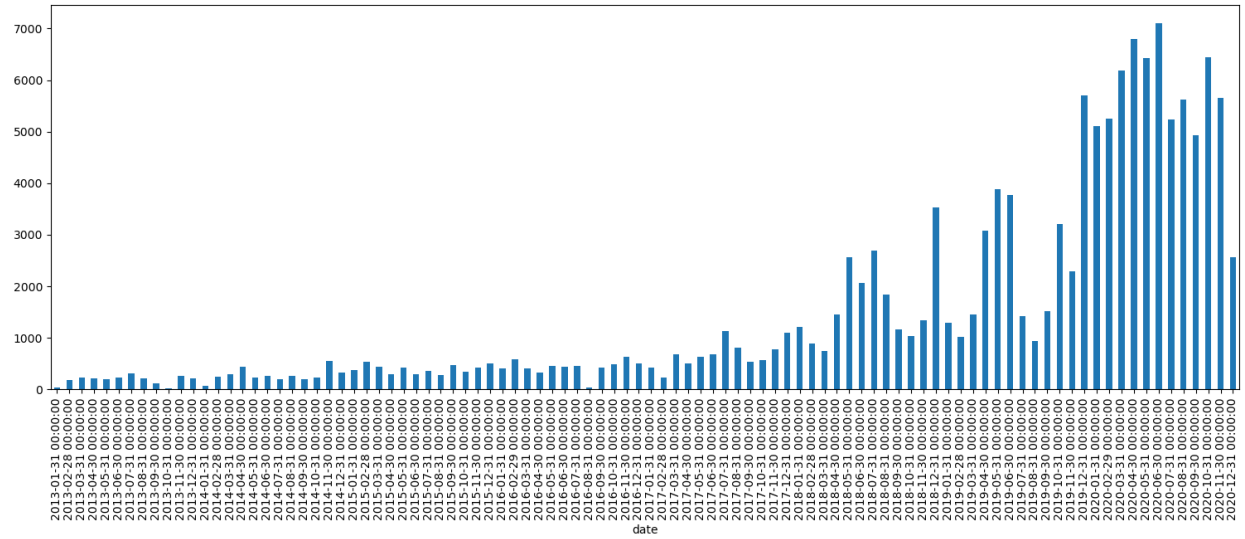
How accurately can entities be identified using natural language processing techniques upon text messages, and how should this accuracy be measured?

Data Acquisition

The data were compiled from the author's own text messages, spanning a time period from February 2013 to December 2020. These text messages were exported from an Android phone in an XML format using a free version of the "SMS Backup & Restore" app found in the Google Play Store in December of 2020. This application exported phone call logs, SMS and MMS data, to include pictures sent via MMS. A total size of 2.9GB, this data export contained 141,155 SMS and MMS messages, of which MMS messages may have multiple parts, which implies the total count is a low estimate of the true number of messages sent and received.

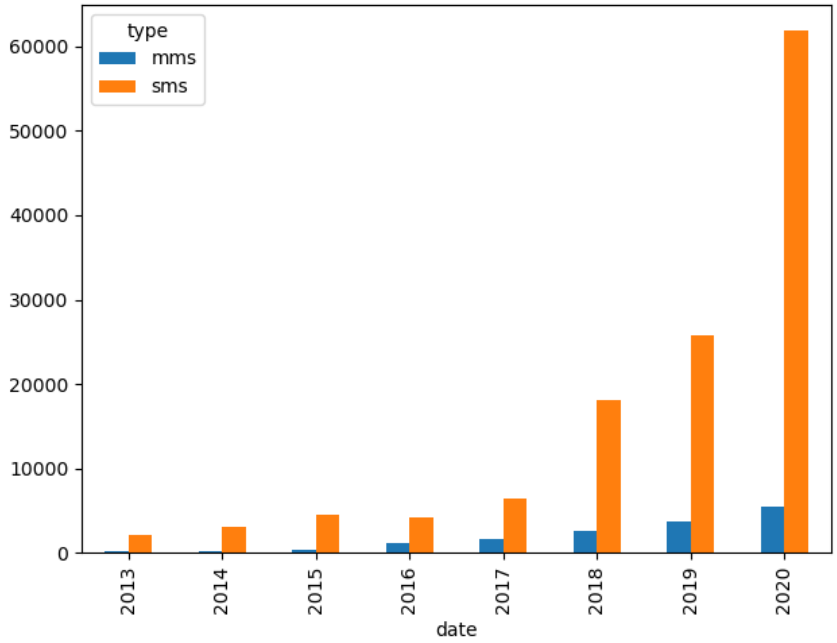
This XML data was then processed using custom python scripts to be in a consistent format usable for further exploration and analysis. The final counts showed 829 unique numbers after normalization, which was associated with 189 unique names. This shows a minimum size to the contacts list, while indicating that many more unlabeled phone numbers were sent or received texts from the author. The following is a graphical description of the amounts and types of messages sent and received over this time.

Text Messages by Month (2013-2020)



Unrelated to this analysis, but text message frequency increases. This is likely due to texting becoming more of a mode of communication for the author, as other messaging services take a back-seat.

Text Messages by Type, by Year



Of interest is the split between MMS and SMS, showing that most texts cannot contain image data and purely contain text.

Data Preprocessing

After acquiring the data from the mobile Android device, a variety of data quality issues had to be addressed. They are summarized below.

Phone number normalization

- Removed non-digits
- Truncated to first 10 digits (removed country-code, of which only USA expected)
- Removed lengths less than 7 (automated text services use shorter numbers)

Name normalization

- Removed indicators: H, X, BB, CMB, EH, Hinge and anything between parentheses and double quotes. Regex was used to parse these from the data, as they were metadata unassociated with person entities
- Removed surrounding whitespace

Text normalization

- Image-only MMS had 'null' text removed
- Merged text messages from each phone number together, creating one large document for each number
- Truncated total text lengths to 1000 characters

Manual Data Labeling

After Spacey initially parsed the sample of text messages to analyze, these messages were exported in a user-editable format to a csv file and imported into Excel. Excel was then used as a simple user-interface to add '1's to indicate the presence of terms that should be recognized as Person entities (and only Person entities). This data was then imported back into the Python code and used as a baseline for determining the various measures of model accuracy upon the text message data.

Model

Spacey was used to perform the steps of tokenization, part-of-speech (POS) tagging, and naming of entities. The model chosen for this was en_core_web_md, a model 48MB in size, trained on blogs, news, and comments from the web. It consists of many thousands of keys, and some smaller number of unique vectors, mapped along 300 dimensions. The claimed accuracy of the NER component of this model received an F1 score of 86.2%. This score includes all entity types. The entity types of particular interest in this analysis are Person entities. As this model was already trained, no cross-validation was performed to further evaluate model performance.

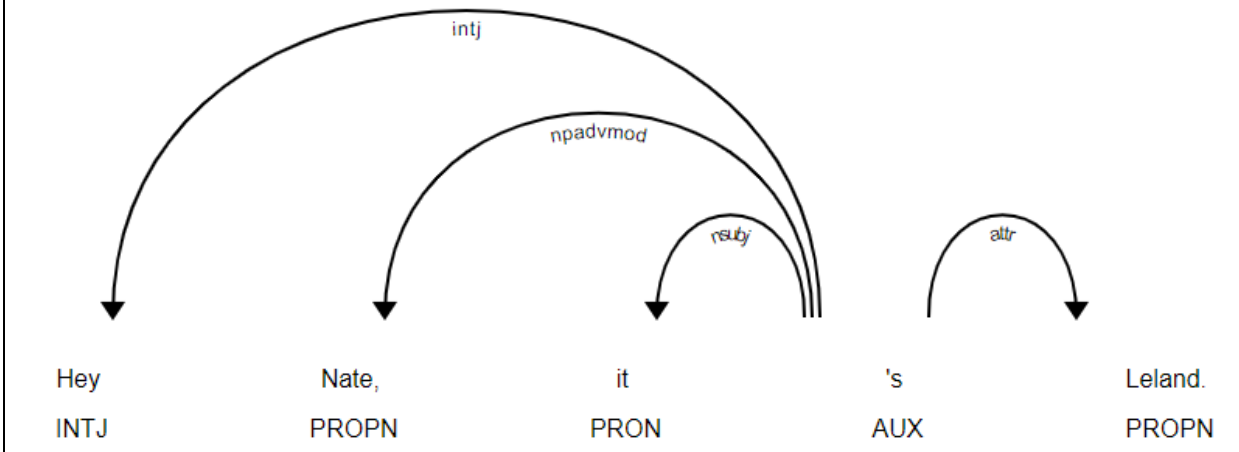
Analysis

The model was run on a randomly sampled subset of 32 sets of text messages, truncated at a length of 1000 characters, which on average included 90 tokenized terms. Measurements for precision, recall, accuracy, and an F1 score was calculated for each set of text messages from each phone number.

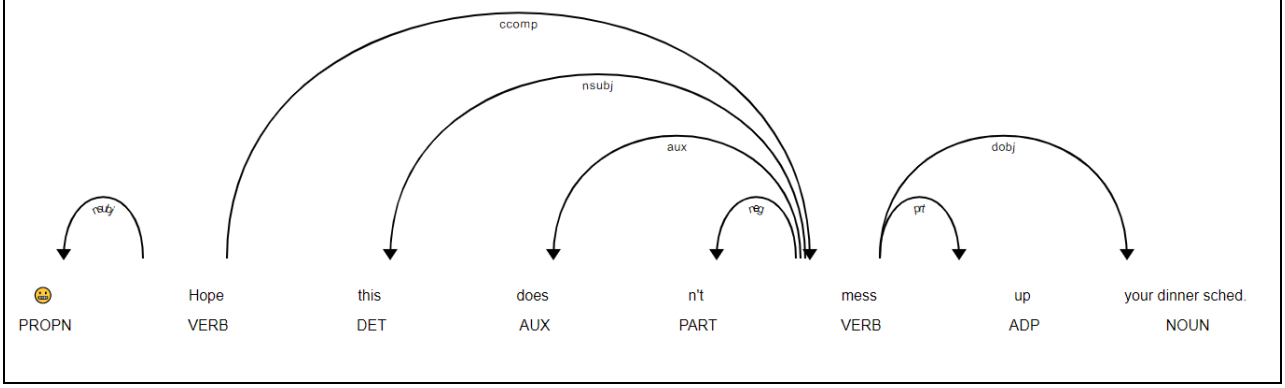
Examples of Model POS Tagging

The Spacey model performed in a less-than-ideal fashion upon this dataset. This is to be expected, since text messages are often of dubious quality, and incorporate symbols and colloquialisms that may not be found in the content the model was trained upon, which was predominantly websites and web comments. The following examples show how in addition to grammatical issues, text communication over mobile phones are prone to other mislabeling. In one case, the identification of an emoticon as a proper noun may result in its classification as a Person entity. This would appear in the confusion matrices as a false positive.

A short, but otherwise well-formed sentence parsed correctly



A sentence with an emoticon parsed as a proper noun

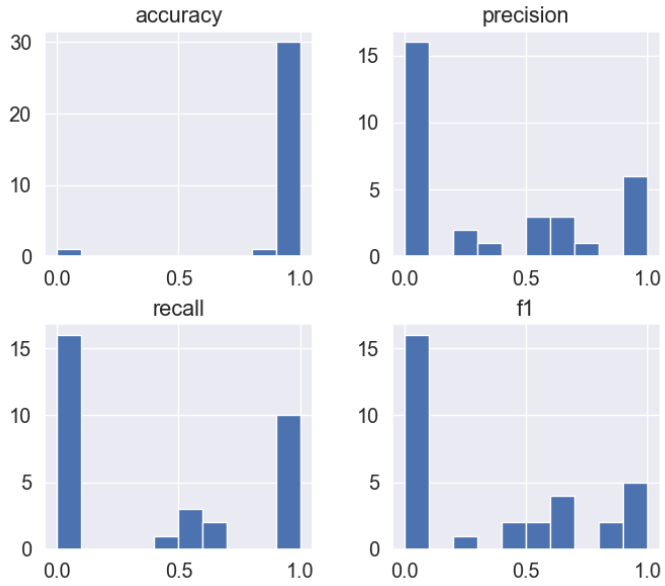


Results

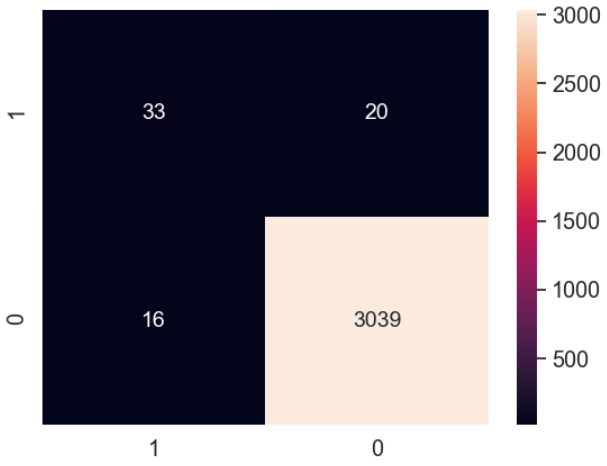
Spacey Name Entity Recognition

Spacey is a library for natural language processing with the ability to carry out tokenization, POS tagging, dependency parsing, and Named Entity Recognition, in addition to much more. This last capability was put to the test against the hand-labeled SMS data. The results are included below.

Plots of each document's accuracy, precision, and recall for identifying Person entities



Confusion Matrix of the Spacey NER algorithm's performance (Left: Predicted, Bottom: Actual)
Note the high accuracy for determining true negatives, but only correctly identified about 62% of actual Person entities.



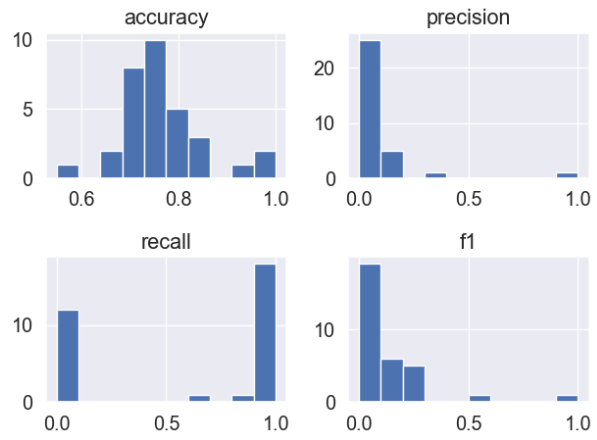
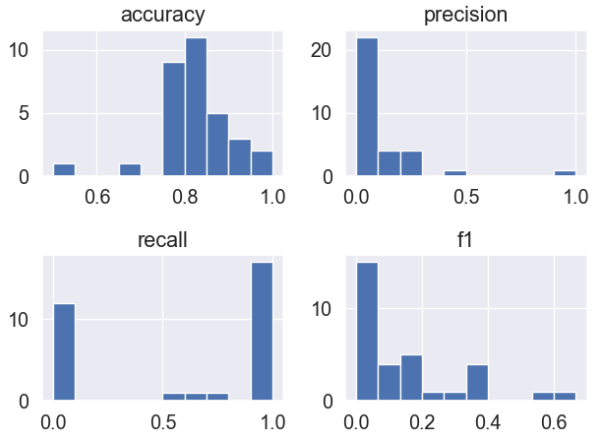
Means of each measurement for all 32 documents:
Accuracy: 0.957820
Precision: 0.344081
Recall: 0.413542
f1: 0.358310

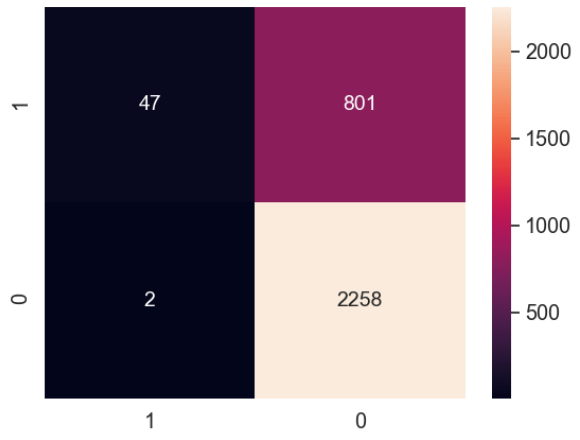
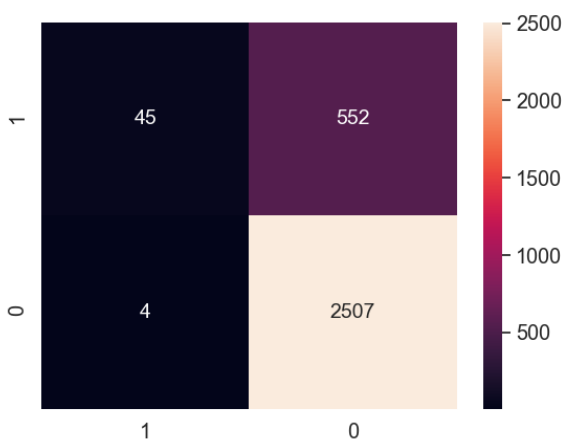
These figures highlight the interplay between accuracy and a good model. While a high degree of accuracy (95% in this case) implies that the model only has a one-in-twenty chance of incorrectly predicting whether a given term is a Person entity or not, this is namely because Person entities are sparsely populated among the tokens of each document. Some documents do not contain Person entities at all. Indeed, guessing that all tokens were not Person entities would result in only a slightly less accurate score. The F1 score does a better job of accounting for the disconnect between accuracy and the model's actual fitness for this task.

The F1 score in this case is 35%, which is a far cry from the high 80% F1 score that the model received when testing on its own corpus of web scrapings. The precision of this model was measured by taking the ratio of true positives to that of the sum of all correctly identified tokens. This measurement increases as it is able to correctly identify Person entities from tokens that aren't. Recall was also calculated to be the ratio of correctly identified Person entities over the sum of both correctly identified entities and false identifications of entities. Recall scores decrease if every token were to be identified as a Person entity. Using these metrics, we see that this model still outperforms a naïve baseline that was carried out next.

Baseline Names Entity Recognition

One kind of baseline considered for identifying Named Entities (specifically those of persons) was to simply compare the tokenized text with a dictionary of given names and surnames. It was hypothesized that this approach would produce a high number of true positives, but also a high number of false positives. The two name lists were acquired and incorporated for testing against the labeled data. [1][2]

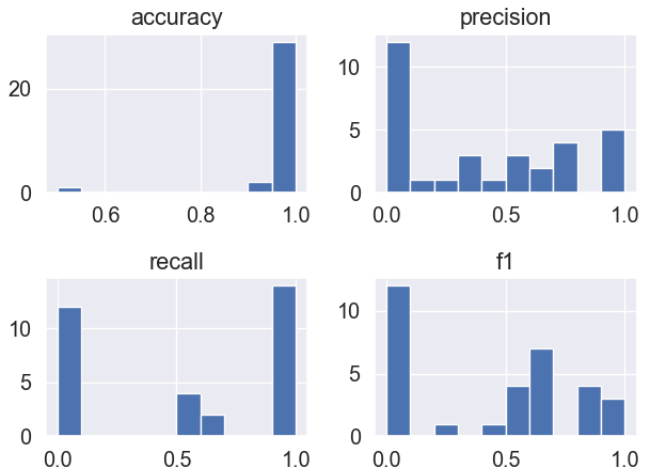
Given and Surnames (over 185,000 names)	Given Names Only (over 99,400 names)
	
Confusion Matrix (Left: Predicted, Bottom: Actual)	Confusion Matrix (Left: Predicted, Bottom: Actual)

	
Means of each measurement for all 32 documents: accuracy 0.767547 precision 0.085021 recall 0.610677 f1 0.124367	Means of each measurement for all 32 documents: accuracy 0.821555 precision 0.103208 recall 0.591146 f1 0.142027

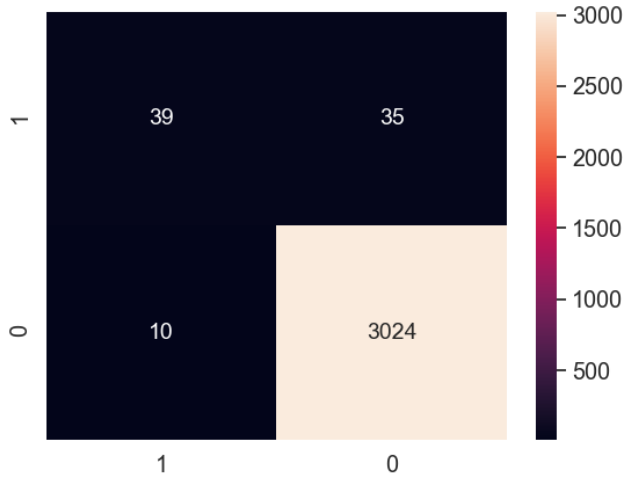
As was hypothesized, this approach performed well when looking at true positives. Almost all of the person entities (all 49 of them) were identified when using both surnames and/or given names. However, low F1 scores resulted from falsely identifying hundreds of tokens as being Person entities. Such entities could be names that are also common dictionary words, like the name “Lane”, also a word for street in English. The abysmal performance of this approach shows that the Spacey model is superior to a naïve approach. It is important to note that recall would not be a sufficient metric alone to judge this baseline approach. While it received a higher recall score than the Spacey model, it is because recall does not take into account false positives (the top-right corner of the confusion matrix).

Names and Dictionaries for Person Entity Recognition

Finally, a last attempt at developing a baseline for comparison of Person Entities was attempted by stripping the English dictionary terms from the names corpus. It was hypothesized that the number of false positives would decrease significantly, since some common names are identical to English words, but most names exist outside of the set of English words. This dictionary of English terms was acquired from the Natural Language Tool Kit’s (NLTK) English dictionary.



Confusion Matrix (Left: Predicted, Bottom: Actual)



Means of each measurement for all 32 documents:

accuracy	0.967847
precision	0.394196
recall	0.543490
f1	0.425967

The results of this baseline analysis still shows more false negatives than the Spacey model, but fewer false positives, and many more correctly identified true positives (actual Person entities) than the Spacey model. In fact, this baseline dictionary+given_name model performs better than the Spacey model by all metrics analyzed: accuracy, precision, recall, and F1 score. As was hypothesized, the number of false positives decreased. In addition, the ability to recognize true positives did not decrease much, meaning that this approach still outperforms the Spacey classifications.

Conclusions

In conclusion, the Spacey model for Named Entity Recognition performs much poorer on SMS conversations (text messages) than a "baseline" simple corpus of names with English dictionary words removed. The Spacey model received an F1 score of 35% versus the baseline's F1 score of 42%. It is

theorized that Spacey underperforms in this area due to the colloquial terms and unpolished nature of the text data it is asked to classify, which often does not contain grammatically correct sentences for which proper nouns can be correctly parsed. As was discovered in this research, it is important to go beyond a simple accuracy metric when profiling the performance of algorithms upon this data, as the entities of interest are only sparsely populated throughout. The F1 metric captures more of the performance characteristics for such models.

References

[1] Given Names List (acquired December, 2020). USA Social Security Popular Baby Names List: https://www.ssa.gov/oact/babynames/limits.html

[2] Surnames List (acquired December, 2020). Erik Norvelle International Surnames Database: https://github.com/smashew/NameDatabases/blob/master/NamesDatabases/surnames/all.txt