Artificial Neural Networks Applied to Named Entity Recognition of Structured Data Sets

Master's Thesis, Submitted on 11/2/2017

Miriam Herman
Yeshiva University - Department of Mathematical Sciences
215 Lexington Ave
New York, NY
miriam.herman@mail.yu.edu

ABSTRACT

This paper demonstrates the utilization of Artificial Neural Networks (ANNs) to classify the contents of columnar data in structured data sets. Using a simple ANN with Glove embeddings, we demonstrated a significant improvement over the existing Stanford NLP, OpenNLP, and NLTK toolkits for the task of classifying names, organizations, and addresses. One of the challenges of creating these sort of classifiers is the problem of explainability. We used LIME to inspect the quality of our models to examine if they were paying attention to the right features. Furthermore, there is the additional issue of how to deal with the fact that negative examples for any classifier constitute an open set, and hence, false positives are a serious problem with these classifiers. We provide an initial demonstration of how we can use fine tuning techniques to change the model if it handles data incorrectly.

Keywords

Artifical Neural Networks; ANN; Named Entity Recognition;NER; structured data

1. INTRODUCTION

According to a survey conducted by CrowdFlower, a platform for data scientists, "data preparation accounts for about 51% of the work of data scientists" and "60% of data scientists view data preparation [collecting, labeling, cleaning, and organizing data] as the least enjoyable part of their work." [1] The report also stated that 49% of a data scientists work involves structured data. It would be extremely valuable therefore to provide a rigorous method for semantically tagging the contents of structured data sets to minimize the amount of time data scientists must spend on the tasks they enjoy the least. Semantically structuring the data is therefore an important first step to allow data scientists to search for useful data, merge data etc.

One mechanism to tag columns semantically involves recognizing the semantic types of entities in a given column. If a column contains entities that are all addresses for instance, it may help the user to find locational data even if the column name were not recognizable as an address. In natural language processing (NLP), the task of extracting semantic types for corresponding words in a sentence is called named entity recognition (NER). There are several NLP models geared for named entity recognition of data. Could we re-use these models to extract columnar types?

There is a significant problem in applying NLP models for named entity recognition on structured data because they have all been trained on unstructured data. The context they use for named entity recognition is therefore significantly different from structured data, which often contains little or no clues for classification of an entity. Our suspicion was that these models would not transfer to well on unstructured datasets because of this fact.

Our goal therefore was to create a set of named entity recognizers uniquely geared towards classifying structured data. We decided to create models using deep learning networks primarily because there are now many language models in deep learning that can be leveraged to build better named entity recognition (e.g., word embeddings that capture the distributional semantics of words from massive corpora). As a baseline, we compared the results of our ANN models to NLP models such as Stanford NLP, openNLP, and Python's NLTK toolkit to find a 99.7% accuracy improvement in the best case for Neural Networks and a 16% improvement in the worst. Compared with all the NLP models, the average improvement was 75% with our Neural Network model. Our source code and training sets can be found at https://github.com/miriamherm/ClientClassification.

Having built the models, we used state of the art techniques to convince ourselves of the validity of the models. We used LIME [5], a tool to inspect what features the classifier was paying attention to. In our example these would be the specific words in the entity that caused the classifier to choose a specific class. We also demonstrate how any classifier is always prone to false positives, mainly because the negative examples for a specific class is really an open set. Given that a classifier can never be perfect, the next question is how do we effectively tune the classifier when it produces false positives. We provide a very initial set of studies around this problem.

DOI: 10.1145/1235

2. DATA, BASELINES, AND EXPERIMENTS

In this section, we describe the datasets we used for training and testing, as well as the results of the baseline classifiers with examples. We also describe the architecture for a Neural Network that achieves near perfect accuracy on the test data set. We then show how to inspect the Neural Networks to assess the quality of what was built, and then present a viable way to fine-tune the model when it makes errors.

2.1 Data Collection and Cleansing

Data for training and testing was acquired from a variety of sources, with duplicates removed. The appendix contains a complete list of data sources used, with notes regarding file name, column name and any additional steps taken on the data.

Once all the data was collected, the data was divided by type and compiled for each type into a single file. The final count includes 723,553 unique addresses, 1,061,544 unique companies, 249,227 unique peoples names and 559,591 unique products. The data sets were then shuffled, and 20% of each set was split for testing purposes.

We also generated a set of 250,000 names using a datafactory [3] that utilizes US census information to create name data determined by name frequency. We performed tests and training on names data with and without the generated names.

2.2 Baselines

In order to assess the necessity of a Neural Network trained classifiers for structured data, three baseline measures were tested; Stanford NLP NER [4], OpenNLP, and Python's NLTK package.

We tested the models on address, company, and names data, using the following metric. A tag was labeled correct if all words in a cell had the same label, and that label was correct, otherwise it was labeled incorrect. If the model allowed, we also show the metric for all words in a cell having at least once the correct label, and the rest "other."

The results of each model are below.

2.2.1 Stanford NLP NER

We used the Stanford NLP NER tool [4] to classify addresses, organizations and people, but because it ships with models for 7 specific types: Location, Person, Organization, Money, Percent, Date, Time, we did not use it to classify product data.

Using the above first metric, we found that Stanford NLP could not classify addresses. It classified items crucial to identifying addresses, such as street numbers and words like "Rd" and "Ave" as "Other". A typical example of classifier output is: 603/O HINMAN/PERSON RD/O where "603" and "RD" were labeled as "Other" and "FINMAN" labeled as "PERSON". For the 144,709 addresses in the test set, the model had 0% accuracy, labeling only two addresses correctly. If we consider the metric of accuracy to be "correct label" and "other" the accuracy on this set increases to 9.28% or 13,431 addresses labeled correctly.

The classifier also had difficulty classifying organizations for a very similar reason; words crucial to an organization, such as "Limited", are classified as "Other". A typical example of the output for an organization cell is: ALASTAIR/PERSON WRAY/PERSON LIMITED/O. Since many companies are named after their founder, they are often confused for names. For the 212,307 companies in the test set, the model had 0.018% accuracy, labeling only 38 companies correctly. If we consider the metric of accuracy to be "correct label" and "other" the accuracy on this set increases to 8.95% or 19,009 companies labeled correctly.

Since the Stanford NLP NER classifier does not distinguish a single letter as a name, and because a lot of our name data consists of first and middle names listed with an initial, it struggled to classify people as well. A typical example of the output for a name is:I/O SONI/PERSON, where the initial is labeled as "Other". Of the 49,844 names in the test set, the model had 3.66% accuracy, labeling 1825 examples correctly. If we consider the metric of accuracy to be "correct label" and "other" the accuracy on this set increases to 52.73% or 26,287 names labeled correctly.

When the Stanford NLP NER classifier was tested on the collected names data together with the datafactory generated names data, it performed significantly better, labeling 39.69% or 35,634 out of 89,789 names correctly. (Missing 2486 generated names.) If we consider the metric of accuracy to be "correct label" and "other" the accuracy on this set increases to 71.89% or 64,548 names labeled correctly.

Out of 406,860 examples, only 1865 or 0.5% of examples were labeled correctly. The results can be found in Table 1.

2.2.2 OpenNLP

We used the OpenNLP NER [2] tool to classify addresses, organizations and people, but like StanfordNLP, it ships capable of classifying: Location, Person, Organization, Money, Percent, Date, and Time. We did not attempt to train it to classify product data.

OpenNLP performed much better than Stanford NLP when classifying addresses because it seems to recognize certain words common to addresses, like "Ave" and "St". However, it does not recognize other crucial words, like "Rd" and misclassified most "Rd"s as "Organizations". Out of 144,709 addresses in the test set, OpenNLP classified 19.22% or 27,818 examples correctly.

This model also performed much better than Stanford NLP when classifying companies, recognizing common company abbreviations like "LTD" and "CO" as part of an organization. However, it failed to recognize common company words like "Limited" and "Agency". Out of 212,307 companies in the test set, OpenNLP classified 36.39% or 77,265 examples correctly.

When classifying names OpenNLP performed significantly better than Stanford NLP because it classified full terms, and didn't classify names with initials if the last name was recognized strongly as a name. A name like "A G QUITO," with an uncommon last name, would have been misclassified as an organization and some names that were not formatted well were left unclassified, such as "B HARRIS". Out of 49,844 names in the test set, OpenNLP classified 63.16% or 31,483 examples correctly.

When the OpenNLP NER classifier was tested on the collected names data together with the datafactory generated names data, it performed slightly better, labeling 70.58% or 63,374 out of 89,789 names correctly.

Out of 406,860 examples OpenNLP classified a total of 136,576 or 33.6% examples correctly. The results can be found in Table 2.

Table 1: Stanford NLP Confusion Matrix

	Address	Company	Name	Product	Other	Mixed
Address	0.000	0.000	0.000	NA	.249	.750
Company	0.000	0.000	.002	NA	.573	.425
Name	0.000	0.000	.037	NA	.140	.824
Product	NA	NA	NA	NA	NA	NA

Table 2: OpenNLP Confusion Matrix

	Address	Company	Name	Product	Unclassified	Mixed
Address	.192	.257	.040	NA	.076	.435
Company	.001	.364	.003	NA	.002	.630
Name	.003	.012	.632	NA	.061	.292
Product	NA	NA	NA	NA	NA	NA

2.2.3 NLTK Classifer

We attempted to classify addresses, organizations and people using the Natural Language Toolkit (NLTK)[7] package in Python. The NLTK package can classify: Organization, Person, Location (for example Mount Everest), Date, Time, Money, Percent, Facility, GPE (Geo-political entities, like city, state/province, country). As before we did not attempt to train it to classify product data.

NLTK performed much better than Stanford NLP when classifying addresses but worse than OpenNLP. It seemed to recognize certain words common to addresses, like “East” and “Indian” but not the crucial words like “St”, “Rd”, and “Ave”. The NLTK classifier misclassified addresses as organizations, people, and other, but very rarely confused them with several mixed labels. Out of 144709 address in the test set, NLTK classified 2.61% or 3,771 examples correctly.

This model performed much better than Stanford NLP and OpenNLP when classifying companies, recognizing common company abbreviations like “LTD” and “CO” as part of an organization. Out of 212,307 companies in the test set, NLTK classified 82.73% or 175,631 examples correctly.

When classifying names NLTK performed worse than Stanford NLP and OpenNLP, and it is unclear why. On occasion NLTK would correctly identify a name with initials as a person, and in other cases it would classify them as organizations, GPEs, or other. There does seem to be a correct trend of classifying names as “Person” if there is a first and middle initial, but this is not consistent. Out of 49,844 names in the test set, OpenNLP classified 3.46% or 1727 examples correctly.

When the NLTK classifier was tested on the collected names data together with the datafactory generated names data, it performed better, labeling 20.43% or 18,346 out of 89,789 names correctly.

Out of 406,860 examples OpenNLP classified a total of 181,129 or 44.52% examples correctly. (This increase in total accuracy is due to NLTK’s proficiency at classifying companies.) The results can be found in Table 3.

2.3 Neural Network Training

A key component of building deep learning networks for NLP tasks is the use of embeddings for words derived from very large corpora. These embeddings capture important distributional semantics of words (e.g. the word *king* is related to *queen*), and are crucial in learning to perform generic NLP tasks. In previous work, we conducted a thorough analysis of embeddings, including; Facebook fastText, word2vec, continuous bag of words dependency based embeddings, character embeddings from the one billion word corpus etc, and their relevance for the task of building deep learning classifiers. (Results for this work can be found in the Embedding - Results section of the appendix) For the classification task at hand, we learned that one of the best pre-trained embeddings for this task are the 100 dimensional GloVe word vectors. Our models therefore were trained using these embeddings.

Our original intent in building the models was to perform a grid search for the parameters of our model. We began our tests with the simplest model and created binary classifiers for each data type. This architectural choice of having many binary classifiers rather than a single multi-class classifier was deliberate. It easily allows extension in the system to add more classifiers for new types as we come across new data. Each model was a simple multilayered perceptron with an embedding layer, with a single hidden layer of 128 nodes to extract features specific to the semantic type. These nodes were connected to a single output node for a binary decision. Each model was optimized with the adam optimizer, and trained with 10 epochs. The four models achieved 97%-99% accuracy. As this was a huge improvement from all the previous baselines, we did not see a need to fine tune the architecture further, although as we will see later, there may be a need to revisit the architectural decisions when fine tuning is needed.

Our models were trained on 2,016,028 unique phrases containing 389,808 unique words. The max number of words in a phrase was set to be 10 (most entity names do not exceed 10 words), so our model was trained on 10-dimensional word vectors. If a phrase had fewer than 10 words, its associated vector was padded with zeros.

2.3.1 Test Set Results

For names, we trained two models, one with datafactory generated names and one without. The model with generated names performed 4% better, from 94% to 98%. Going forward in this paper we used the models trained with the generated names when we refer to the name models. The address classifier hovered around 99%, the companies classifier around 98% and the product classifier at 96%. Complete results for all the models are presented in Table 4 and Table 5.

2.4 Business Set Results

Table 3: NLTK Confusion Matrix

	Address	Company	Name	Product	Other	Mixed
Address	.026	.208	.312	NA	.449	.005
Company	.003	.827	.005	NA	.094	.070
Name	.002	.145	.035	NA	.792	.026
Product	NA	NA	NA	NA	NA	NA

Table 4: Generated Names NN Confusion Matrix

	Address	Company	Name	Product
Address	.997448	.0017542	.0003253	.003252
Company	.00128143	.987642	.002799	.007761
Name	.0010614	.008826	.989925	.01163
Product	.00407549	.0178114	.006845	.965432

Table 5: Collected Names NN Confusion Matrix

	Address	Company	Name	Product
Address	.996454	.001618	.000358	.00376
Company	.000656	.987531	.002567	.007580
Name	.006821	.0619743	.949441	.0271665
Product	.00291918	.0184617	.0068049	.965063

In order to validate the accuracy of the above Neural Networks results, we tested the models on a generated business data set kept completely separate from the training process. This step simulates the effect of trying these binary classifiers on a new set of columnar data (see Table 6 for example columns in this dataset). As shown in Table 6 the baseline ANN performed well in correctly identifying people's names, and address details, but was not as confident about company classifications in the data. It also seemed to miss components of addresses such as cities or states when they appeared without their street addresses. The most egregious examples of incorrect classification though had to do with false positives. As an example, dates were mislabeled as products 99% of the time, and column entries with single letter codes for "pay cycle" were classified as people's names 98% of the time. All results for the business set appear in the table as "Baseline ANN" results, and focus on the first set of results we obtained when we tried generalizing the built in classifiers to new unseen data types.

It was apparent that the binary classifiers we had built, despite their excellent test performance had some serious problems in their modeling. We tried to address problems in the modeling in two ways:

- We used a tool called LIME to understand exactly what words the classifier was paying attention to.
- We tried to see if using other network architectures such as LSTM (Long Short Term Memory [6]) might help alleviate the problem. The core idea behind LSTM is that it is used for capturing sequences of information. An initial by itself is not likely to be a name but an initial embedded with a first name and a last name is likely to be that of a person. Similarly, a set of numbers in a date should not trigger a classification of product because there are no other product terms associated with the numbers. Using an LSTM might reduce false positives, by helping build in context. We tested this hypothesis.
- We tried to address the problem of false positives on unseen data types. As we described earlier, classifiers are always prone to false positives, primarily because it is actually never possible to show them all possible negative examples at training. When a classifier produces a very high rate of false positives we have one of two options: either re-train the original model using a sample of the false positives as negative examples, or fine tune the weights of the existing model to retain the true positives as best as possible while reducing the false positive rate. We examined the role of each in a preliminary study.

Each of these approaches is described in the section below.

3. TROUBLESHOOTING THE CLASSIFIERS

3.1 LIME - Explaining the Predictions

Machine learning algorithms, and deep learning systems in particular, are black box solutions, where input is fed to a model and a classification returned without an explanation as to how the decision was made. LIME, (Local Interpretable Model-Agnostic Explanations) [5], attempts to solve this problem by learning an interpretable model locally, around the prediction, to provide insights into exactly what the model might be doing. Using LIME on a classifier, one can discern which words were most crucial to the classification process. This sort of exercise is helpful in understanding whether the classifiers are actually useful or whether they picked up spurious correlations in the sample.

We provided LIME with 100 words from each type and list in Table 7 the top 6 results. A complete list of results can be found in the appendix [1].

As one might expect, for an address classification model 'St' is among the 6 most important words. For the company classifier we see words such as 'Limited' as one might expect, but surprisingly single letters 'W', 'A', and the irregular word 'Bowker' at the top of the list. We observe for names a list of common names. For Products, LIME provides some insight into why dates were being mislabeled as products, as we describe in the next section. Numbers, like '32' and mixtures of letters and numbers, like "pack, 1" and "1X10", carry a lot of weight in the classification of a product. This explains the spurious results we saw with dates.

3.2 LSTM

We changed the network architecture to feed the embeddings for the same 10 dimensional word vectors to 128 LTSM

[1] Our attempts to provide the entire training set to LIME proved to be problematic because the system could not handle the sample set sizes on our machines

Table 6: Business Data Results

Col Name	Example	Baseline NN	LSTM 5 Epoch
Account Owner	Toni Gomez	Person - 0.998	Person - 0.999
Billing Address	Suite # 10049	Address- .99 4	Address - .69
Billing Contact	Allen Hardin	Person- .999	Person- .999
Billing Email	jgoff@ma1l2u.org	Company - .826	Company - .956
City	Helena	Person - .429	Person .356
Conversion Date	Thu Apr 03 05:20:32 EDT 2014	Product - .999	Product - .998
Country	Slovakia	Product - .430	Product - .413
Custom Metrics	code,text,room	Product - .584	Product - .564
Org Name	Morgan Studios	Company - .864	Company - .877
Parent Name	'Ringgold Cafe'	Company - .861	Company - .852
Pay Cycle	'Q','T'	Person - .982	Person - .997
Pay Method	'Invoice'	Product - .47	Product - .542
po_num	PO6793946273	Company - .575	Product - .397
State	'in', 'so'	Product - .664	Product - .880
Street	'Mcconnel'	Person - .501	Person - .456
Terms	'Standard'	Product - .661	Product - .552
Valid From	Thu Jul 30 05:20:32 EDT 2013	Product - .999	Product - .999
Valid To	Sat Mar 15 05:20:32 EDT 2014	Product - .999	Product - .998

Table 7: LIME - Most Important Words

Data Type	6 Important Words
Address	Crse, Rock, St, Middle, Pendleton, Attawanhood
Company	LIMITED, BOWKER,W, A, LIMITED IC, CROWTON
Name	SAN, JOSE, DELA, BUCHANAN, BONNER, HICKMAN
Product	Hips, 32, Support, Cheese, 3, lb

Table 8: LSTM 5 epochs - Confusion Matrix

	Address	Company	Name	Product
Address	.998772	.000653	.000198	.000594
Company	.000338	.99354	.00206	.00362
Name	.000678	.00407	.990576	.00397
Product	.001436	.00768	.00551	.98511

nodes, with dropout and recurrent dropout of .2, and sigmoid activation. We trained for 5 epochs. It took more than 10X longer to run a single epoch than the previous model took for 10 epochs. Testing it took significantly longer as well. Results are in Table 8. As shown in the Table 8, the base classification performance of an LSTM model is excellent as one might expect. Unfortunately, as shown in Table 6, this architecture by itself did not help reduce the false positives to unseen data types. It appears that even when the new data types did not contain some of the contextual words that must have occurred in the original training set, the model tended to falsely classify dates as products or letter codes as people's names. The lime results (See table 9) for names very obviously reveal this phenomena, with 2/3 of the most important name words for this model being single letters.

3.3 Retraining the model versus fine tuning the model

3.3.1 *Retraining the model with dates*

We added 8487 dates as negative examples and retrained the model on 10 epochs using our initial architecture. The retrained model displayed slightly improved results from the original, with accuracy ranging from 96.6%-99.6%, compared to 94%-98%. On the business data set, this model and the baseline NN model classified every column the same way, but this model did not classify dates as products, companies, addresses, or names. This model also stood slightly apart from the initial model on Pay Method and Terms, classifying both columns with more confidence as products, from 47.2% confidence to 78.8% for Pay Method (Invoice, Strip) and from 66.1% to 79.5% for Terms (Standard). It also stood out for Billing Address with a decrease in confidence from 99.9% to 85.4%. These results can be found in table 14. The results of this model on the test set can be found in table 10.

The first 6 lime results of this model can be found in table 11. Our intention when retraining this model with dates data was to improve the products model, and with that we were successful: no words containing letters and numbers are in the top 6.

3.3.2 *Retraining the model with letter codes*

In order to prevent the model against classifying single letters as names we added the 26 letters in the English Language to the set of negative examples and retrained the model on 10 epochs using our initial architecture. With this method we were unsuccessful. The limited number of negative examples barely improved the model, with single letters in the business data set being classified as names with 95.6% confidence. For complete results of this model on the business set, see table 14. See table 12 for complete results of this model on the test set.

Table 9: LIME - Most Important Words - LSTM

Data Type	6 Important Words
Address	Rd, Jarvis, Three, 33,Fish, Milford
Company	Limited, Properties, Limited I.C, Blueberry, Property, Crowton
Name	Virella, Coffey, D, K, M, L
Product	Blender, Bar, Moisturizing, Waring, Mushroom, Hips

Table 10: Network w/ Dates- Confusion Matrix

	Address	Company	Name	Product
Address	.996	.001	.0004	.004
Company	.0008	.987	.002	.008
Name	.001	.009	.987	.011
Product	.002	.018	.006	.966

The first 6 lime results of this model can be found in table 11. The model was unsuccessful and as we can see from the lime results, single letters are still highly related to names.

3.4 Fine-Tuning

Our original model performed quite well on the data types it had seen before, such as names and companies, but as mentioned above, was confused by some new data types like dates. Instead of retraining the entire model with the new data types, we tried fine tune the model with 8487 dates of different formats. Fine tuning is especially attractive in actual deployment, because the model can be tweaked as it encounters more and more negative examples that are confusable with positive ones.

As per the standard recommendations [8] We fine-tuned the model using Stochastic Gradient Descent. Freezing all layers except the connection to the output layer, we set the learning rate to 0.0001, decay to 1e-6 and momentum to 0.9.

We compiled and trained each model with the new optimizer on 50 epochs, using only the negative examples.

3.4.1 ANN - Products

When fine-tuning the baseline NN, the run time was less than a minute and accuracy on the product test set dropped from 96.54% to 92.29% . The model initially classified dates as products with 89.20% confidence, and after the fine tuning, that confidence drops to 4.40%. A good start, but further testing must be done on the business set.

When we ran the model on the business set we found that dates are classified as products with only 33.51%-43.10% confidence, way down from the original 99.9%. We achieved our goal, fine tuning the model to not mistake dates for products, but in the process lowered the classifiers overall accuracy.

If we could improve the accuracy on the test set higher than 92.29%, with the same results for dates, that would be ideal.

We lowered the number of epochs in half, to 25, and found a 1% improvement to 93.44% accuracy on the test set, but dates are classified as products in the business set with 74.60%-79.41% confidence. We lowered the decay to 1e-4 and the accuracy on the test set increases less than 1% to 92.54% but dates are classified as products with 43.30-50.58% accuracy.

We tried fine-tuning the LSTM model trained with 5 epochs, and accuracy on the product test set barely dropped from 98.51% to 98.43% . The model initially classified dates as products with 86.88% confidence, and after the fine tuning, that confidence barely dropped again, to 84.28%.

There is a very clear trade off here between false positives and false negatives, and it appears that we need more rigorous methods for fine tuning the weights of a model. This is clearly important for future work.

3.4.2 ANN - Names

Using the same architecture as used above for dates we tried to fine-tune the model by providing each of the 26 letters in the English language as negative examples. Our goal was for the model to not classify single letters as names.

Initially the model classified single letters as names with 90.5% confidence, and after 50 epochs the confidence dropped to 88.6% confidence. The confidence for the valid test data dropped .1% from 98.8% to 98.7%.

When the number of epochs was increased from 50 to 500, the model improved, classifying letters as names with only 47.0% confidence. The confidence for the valid test data dropped too, from 98.8% to 95.0%.

When we performed the fine-tuning on the LSTM model trained on 5 epochs the accuracy on the test set decreased from 99.1% to 98.9% and accuracy on the letters set stayed exactly the same, at 99.2%.

There remains a trade off between false positives and false negatives, and a more rigorous fine-tuning method is needed.

4. OPEN ISSUES: MISSING WORDS

Was the creation of classifiers with word embeddings a problem in our classification task? For the 389917 unique tokens found by the tokenizer, a total of 145,573 words (37%) were excluded from the embedding matrix; 8724 of 21,711 address words (40%), 35985 of 252378 company words (14%), 33669 of 66475 names words (51%), and 67195 of 106293 product words (63%). About 6% of the total number of texts.

Of the missing address words, the vast majority are numbers, including 981764 and 4615 and numbers with words like “u52” and “569a”. However LIME did list numbers very highly in its results for an addresses most important words.

Of the missing company words, the vast majority are non-English words, including “kebabse”, “capitus”, “amav”, and “bevtex”. Some words like “limitmanagement” were also discarded. Though it would seem most non-English words were excluded, LIME did list very unlikely words like “Bowker” in its results for most important words.

Of the missing name words, every single one is an irregular name such as “modrcin”, “repohl”, and “berdichevskaya”.

Of the missing product words, many are numbers, dimensions, or non English words like “colgadores”, however there are some description words like “fisherman’s” and “minilight” that are being ignored that may be causing the product clas-

Table 11: LIME - Most Important Words - Dates

Data Type	6 Important Words
Address	Saddle, Ave, Rd, Suffield, Madison, St
Company	Limited, W, A, Properties, Bowker, Innovations
Name	San, Barker, W, Coffey, Matthew, Stephanie
Product	fluid, Surimi, Protect-a-Bed, Mixed, Greek, L-lysine

Table 12: Network w/ Letters- Confusion Matrix

	Address	Company	Name	Product
Company	.0007	.988	.003	.008
Name	.0008	.011	.985	.010
Product	.003	.020	.006	.966

Table 13: LIME - Most Important Words - Letters

Data Type	6 Important Words
Address	Three, Fish, Mile, Colony, Trl, Rd:q
Company	Limited, Ltd, Properties, W, A, Blueberry
Name	Theresa, S, A, Jose, Glen,K
Product	60, Traditional, Free,Travel, Moisturizing, Skin

sification difficulty.

One might argue that word embeddings for type classification is actually useful because it focuses the learning problem on what are clearly generalities rather than having the network learn some idiosyncratic words. However, the dominance of odd words like 'Bowker" or "2B" in the embeddings may have skewed the network. A character embedding might yield better results for classification but this is future work.

5. CONCLUSIONS

In conclusion we have shown the viability of Neural Networks for learning the contents of structured data sets, and have proven that this method is an improvement on the existing software. We have also presented a simple usable architecture for this task, as well as tools for fine-tuning a model with a small set of negative examples. The source code for this project can be found at https://github.com/miriamherm/ClientClassification along with our training and test datasets.

6. ACKNOWLEDGMENTS

I would like to acknowledge my advisor, Dr Kavitha Srinivas, for the time and dedication she spent making sure my project was the best it could be. As well I would like to thank the Yeshiva University Department of Mathematical Sciences for their program in mathematics.

7. REFERENCES

[1] 2017 data scientist report. crowdflower.com, 2017.

[2] A. O. D. Community. Apache opennlp developer documentation. https://opennlp.apache.org/docs/1.8.2/manual/opennlp.html, 2011,2017.

[3] A. Gibson. Generate test data with datafactory. http://www.andygibson.net/blog/tag/datafactory/, 2011.

[4] T. G. Jenny Rose Finkel and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL '05 Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. ACM, June 2005.

[5] S. S. Marco Tulio Ribeiro and C. Guestrin. âĂIJwhy should i trust you?
âĂİ, explaining the predictions of any classifier. In *KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, August 2016.

[6] J. S. Sepp Hocreiter. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[7] E. K. Steven Bird and E. Loper. Natural language processing with python- analyzing text with the natural language toolkit. http://www.nltk.org/book/.

[8] F. Xu. A comprehensive guide to fine=tuning deep learning models in keras (part i). https://flyyufelix.github.io/2016/10/03/fine-tuning-in-keras-part1.html, 2016.

APPENDIX

A. DATA SOURCES

B. LIME FULL RESULTS

C. LIME FULL RESULTS

D. EMBEDDING - RESULTS

Table 14: Business Data Retraining Results

Col Name	Example	Network w/dates	Network w/ letters
Account Owner	Toni Gomez	Person - 0.997	Person - 0.991
Billing Address	Suite # 10049	Address- .854	Address- .580
Billing Contact	Allen Hardin	Person- .998	Person- .997
Billing Email	jgoff@ma1l2u.org	Company - .826	Person - .421
City	Helena	Person - .401	Person - .421
Conversion Date	Thu Apr 03 05:20:32 EDT 2014	None	Product - .999
Country	Slovakia	Product - .427	Product - .408
Custom Metrics	code,text,room	Product - .593	Product - .592
Org Name	Morgan Studios	Company - .839	Company - .869
Parent Name	'Ringgold Cafe'	Company - .833	Company - .824
Pay Cycle	'Q','T'	Person - .993	Person - .948
Pay Method	'Invoice'	Product - .788	Product - .493
po_num	PO6793946273	Company - .568	Company - .564
State	'in', 'so'	Product - .576	Product - .629
Street	'Mcconnel'	Person - .438	Person - .457
Terms	'Standard'	Product - .795	Product - .922
Valid From	Thu Jul 30 05:20:32 EDT 2013	None	Product - .999
Valid To	Sat Mar 15 05:20:32 EDT 2014	None	Product - .999

Table 15: Data Sources

Source	Data Type	# records	Notes
open address	Addresses	707094	Concatenated number and street from US Northeast
sec.gov /rules/other	Companies	951	Removed headers
quality frozen foods	Products	5615	Downloads
crown products	Products	5615	Downloads
ikea.com	Products	2765	copied all products from site catalog
product-open-data.com/	Products	551953	GTIN table GTIN_NM column
product-open-data.com/	Companies	4151	brand table brand_NM column
wordlab	Companies	4924	company-names-list Removed top 66 and bottom 33 rows
wikipedia	Companies	688	List_of_common _carrier_freight _railroads_in _the_United_States
wikipedi	Companies	1648	List_of_companies _of_the_United_States
wikipedia	Companies	103	List_of_department _stores_of_the _United_States
wikipedia	Companies	112	List_of_independent _bookstores _in_the_United_States all listed with"in city"
wikipedia	Companies	404	List_of_supermarket _chains _in_the_United_States split on '(' and '-'
wikipedia	Companies	66	List_of_United_States _clock_companies
wikipedia	Companies	259	List_of_United_States _insurance_companies
wikipedi	Companies	502	List_of_United_States _water_companies
census.gov	People	5494	1990_census_namefiles Names generated with most common first names concatenated with most common last names
data.gov	Companies	1139	Active_Benefit _Companies Business Name
data.gov	People	237260	Civil_List Name
data.gov	Companies	4362	Consumer_Complaints Company
data.gov	Products	18	Consumer_Complaints Product
data.gov	Addresses	2112	FOIL_Report - Trade_Waste_All _Approved_or_Denied Mailing Office

Table 16: Data Sources part 2

Source	Data Type	# records	Notes
data.gov	Companies	2151	FOIL_Report - Trade_Waste_All _Approved_or_Denied - Trade Name
data.gov	Addresses	369	IDOL_2013 _Registered_Owner _Ridesċsv Address
data.gov	Companies	1100	IDOL_2013 _Registered_Owner _Ridesċsv Ownername and Manufacturer
data.gov	People	367	IDOL_2013 _Registered_Owner _Ridesċsv ContactName
data.gov	Products	1731	IDOL_2013 _Registered_Owner _Ridesċsv Ridename
data.gov	Companies	1504	Licensed _Insurance _Companies Company Name
data.gov	Addresses	488	Lobbying _Reporting _System _Maryland _Registered _Employers _List _ Address
data.gov	Companies	444	Lobbying _Reporting _System _Maryland _Registered _Employers _List _ Firm Name
data.gov	People	652	Lobbying _Reporting _System _Maryland _Registered _Employers _List _ Concatenate(First Name, Middle Name, Last Name)
data.gov	Companies	184	Lobbyist _Activity _Contacts Lobbyist Firm
data.gov	Addresses	270	Lobbyist _Activity _Contacts_ Lobbyist
data.gov	Addresses	5031	M_WBE_LBE _and_EBE _Certified_Business_List-Address1
data.gov	People	5327	M_WBE_LBE _and_EBE _Certified_Business_List-Contact_Name
data.gov	Companies	5410	M_WBE_LBE _and_EBE _Certified_Business_List-Vendor_Formal _Name
data.gov	Addresses	200	Neighborhood _and_Rural _Preservation _Companies _Directory Street Address

Table 17: Data Sources part 3

Source	Data Type	# records	Notes
data.gov	Companies	205	Neighborhood _and_Rural _Preservation _Companies _Directory Organization Name
data.gov	Products	2306	nsn _extract _4518-Common_Name
data.gov	Addresses	7320	Oregon _Consumer _Complaints Address 1
data.gov	Addresses	1038	Oregon _Consumer _Complaints Address 2
data.gov	Companies	9403	Oregon _Consumer _Complaints
data.gov	Companies	38	OWEB _Small _Grant _Teams Team
data.gov	People	38	OWEB _Small _Grant _Teams Team Contact
data.gov	Companies	1798	Prequalified _Firms - Prequalified Vendor Name
data.gov	Companies	266	SCA _Disqualified _Firms Vendor Name
data.gov	Companies	47	Science _Festival _Company _Sponsors Company Sponsor
data.gov	Companies	58	Top _50 _Employers _ _Hawaii _County - Name
data.gov	People	54	Top _50 _Employers _ _Hawaii _County Concatenate (Contact First Name, Contact Last Name)
data.gov	Addresses	135	Top _Manufacturing _Companies _in _SSMA _Region Primary Address
data.gov	Companies	202	Top _Manufacturing _Companies _in _SSMA _Region Company Name and Ultimate Parent
data.gov	People	233	Top _Manufacturing _Companies _in _SSMA _Region First Name and Last Name
data.gov	Addresses	71	Trade _Waste _Broker _Registrants Address
data.gov	Companies	71	Trade _Waste _Broker _Registrants Account Name
DBpedia	Companies	82838	Column B
DBpedia	People	200946	
data.gov.uk	Companies	1,045,333	BasicCompanyData

Table 18: LIME Results - Addresses (1)

Baseline NN	LSTM	w/ Dates	w/ Letters
’Crse ’ 0.012	’Rd ’ -0.247	’Three’ 0.016	’Three’ 0.013
’Rock’ 0.006	’RD ’ -0.224	’Colony’ 0.013	’Fish’ 0.012
’St ’ 0.006	’JARVIS’ 0.165	’Crse ’ 0.012	’Mile’ 0.011
’ST ’ -0.006	’Three’ 0.133	’Mile’ 0.012	’Colony’ 0.011
’Middle’ -0.006	’33’ 0.122	’40’ -0.009	’Trl ’ 0.010
’Pendleton’ -0.006	’Fish’ 0.115	’St ’ 0.008	’RD ’ -0.010
’Attawanhood’ -0.005	’MILFORD’ 0.098	’Worthington’ 0.008	’Rd ’ -0.009
’Fish’ 0.005	’86’ -0.096	’339’ 0.008	’Rock’ 0.007
’WALNUT’ 0.005	’S’ 0.095	’Hawthorne’ 0.008	’Settlement’ 0.007
’329’ -0.005	’Colony’ 0.087	’Bennetts’ 0.008	’Attawanhood’ -0.007
’Colony’ 0.005	’ST ’ -0.083	’ST ’ -0.008	’Cook’ -0.007
’Meadow’ -0.005	’Crse ’ -0.079	’7’ -0.008	’LN ’ 0.006
’200’ 0.005	’St ’ 0.065	’329’ -0.008	’555’ 0.006
’143’ 0.004	’40’ -0.061	’Fish’ 0.007	’72’ 0.006
’Old’ 0.004	’Dr ’ -0.052	’LN ’ 0.007	’143’ 0.006
’Settlement’ 0.004	’14’ 0.047	’CIRCLE ’ -0.007	’Shadow’ 0.006
’US’ 0.004	’Overlook’ -0.044	’Sunny’ 0.007	’113’ 0.006
’555’ -0.004	’N’ -0.042	’Land’ -0.007	’25’ 0.006
’23’ -0.004	’Cook’ 0.040	’ST’ -0.006	’Crse ’ 0.005
’72’ -0.004	’W ’ 0.039	’S ’ -0.006	’480’ -0.005
’135’ 0.004	’S ’ 0.039	’14’ -0.006	’Pl ’ 0.005
’Trl ’ 0.004	’308’ -0.038	’Morea’ 0.006	’Gail’ -0.005
’Arrowhead’ 0.004	’76’ -0.037	’23’ 0.006	’JARVIS’ 0.005
’Rdg ’ -0.004	’Trl ’ -0.037	’86’ 0.006	’135’ 0.005
’29’ 0.004	’89’ 0.036	’87’ -0.006	’Canterbury’ 0.005
’MILFORD’ 0.004	’329’ 0.036	’45’ 0.006	’308’ -0.005
’180’ -0.004	’59’ 0.035	’RD ’ -0.006	’SOUTHMAYD’ 0.005
’9’ 0.004	’Old’ -0.035	’Farm’ -0.006	’N’ 0.005
’Overlook’ -0.004	’Dziok’ 0.034	’Trl ’ 0.006	’34’ -0.005
’259’ 0.004	’Spgs ’ 0.034	’Pl ’ 0.006	’S’ 0.005
’Wells’ -0.004	’Ohio ’ 0.033	’135’ 0.006	’Sperry’ 0.005
’Highland’ 0.004	’135’ -0.033	’1294’ -0.006	’Weed’ -0.005
’Elm’ 0.004	’472’ -0.032	’Nooks’ -0.006	’Wells’ -0.005
’89’ -0.003	’Morea’ -0.032	’US’ 0.006	’215’ 0.005
’Wood’ -0.003	’MIDDLE’ 0.031	’S’ 0.006	’Sheridan’ -0.005
’Neptune’ -0.003	’46’ -0.031	’9’ -0.006	’KING’ -0.004
’Hawthorne’ -0.003	’471’ 0.030	’512’ 0.005	’67’ -0.004
’TERR ’ 0.003	’Tpke ’ 0.030	’472’ -0.005	’MAIN’ -0.004
’Overhill’ -0.003	’MACAULEY’ -0.029	’32’ 0.005	’AVE ’ -0.004
’CIRCLE ’ -0.003	’South ’ -0.029	’25’ -0.005	’Farmington’ 0.004
’River’ -0.003	’WHITE’ -0.029	’Holcomb’ 0.005	’View’ 0.004
’Oates’ 0.003	’8516’ 0.028	’281’ 0.005	’Naugatuck’ 0.004
’Naugatuck’ 0.003	’2768 ’ -0.028	’Long’ 0.005	’Elm’ -0.004
’Three’ 0.003	’CIRCLE ’ 0.027	’308’ -0.005	’MILFORD’ 0.004
’215’ -0.003	’Arrowhead’ -0.027	’Portland’ 0.005	’JANET’ -0.004
’WACONA’ -0.003	’School’ 0.027	’Highland’ -0.005	’P.O.BOX’ 0.004
’Ct ’ 0.003	’Shadow’ 0.027	’Hudson’ 0.005	’St ’ 0.004
’JARVIS’ 0.003	’Mile’ 0.027	’Sheridan’ 0.005	’180’ -0.004
’750’ 0.003	’KING’ 0.026	’Wells’ 0.005	’Ford’ 0.004
’Randolph’ 0.003	’Ct ’ 0.025	’TPKE’ -0.005	’Pembroke’ -0.004

Table 19: LIME Results - Addresses(2)

Baseline NN	LSTM	w/ Dates	w/ Letters
’37’ 0.003	’Cheshire’ 0.025	’71’ -0.005	’Nooks’ -0.004
’S’ 0.003	’26’ -0.025	’8516’ 0.005	’59’ 0.004
’42’ -0.003	’Myrtle’ -0.024	’Neptune’ -0.005	’Sixth’ -0.004
’Langford’ 0.003	’Attawanhood’ 0.024	’215’ -0.004	’48’ 0.004
’195’ -0.003	’Waterfront’ 0.024	’83’ 0.004	’2410’ -0.004
’1505’ 0.003	’US’ 0.024	’178’ 0.004	’Pendleton’ 0.004
’View’ 0.003	’House’ 0.024	’Rock’ 0.004	’42’ 0.003
’PRATT’ 0.003	’Meadow’ 0.023	’42’ -0.004	’MIDDLE’ 0.003
’Sixth’ -0.003	’Bennetts’ -0.023	’46’ -0.004	’Middle’ -0.003
’7’ -0.003	’High’ -0.022	’Shadow’ 0.004	’WILSON ’ -0.003
’Bennetts’ -0.003	’1173’ -0.022	’Carol’ 0.004	’6’ -0.003
’KING’ -0.003	’Gertrude’ 0.022	’19’ 0.004	’Carol’ -0.003
’Broadway’ 0.003	’Cedar’ 0.022	’Overlook’ -0.004	’Rdg ’ 0.003
’Hilltop’ -0.003	’195’ -0.021	’44’ 0.004	’South ’ -0.003
’2’ -0.003	’View’ 0.021	’AVE ’ -0.004	’Waterfront’ 0.003
’308’ 0.003	’Settlement’ -0.021	’72’ -0.004	’94’ 0.003
’18’ -0.003	’150’ -0.021	’Mountain’ -0.004	’40’ 0.003
’Farmington’ -0.003	’Hawthorne’ 0.021	’Ave ’ -0.004	’12’ 0.003
’14’ -0.003	’47’ 0.021	’Arrowhead’ -0.004	’87’ -0.003
’280’ 0.003	’Soundview’ 0.020	’Pendleton’ -0.004	’2768 ’ -0.003
’WHITE’ 0.003	’Overhill’ -0.020	’28’ -0.004	’150’ 0.003
’Lakeside’ -0.003	’1’ 0.020	’26’ -0.004	’Highland’ 0.003
’F.D.’ 0.002	’Neptune’ -0.020	’30’ -0.004	’Bennetts’ -0.003
’S ’ -0.002	’12’ 0.020	’581’ 0.004	’MACAULEY’ -0.003
’Dr ’ 0.002	’Tree’ 0.020	’82’ -0.004	’House’ 0.003
’Hartford’ -0.002	’73’ -0.020	’ALYCE’ -0.004	’Farm’ -0.003
’581’ 0.002	’87’ -0.020	’PRATT’ 0.004	’ST ’ -0.003
’Porterbrook’ -0.002	’Russell’ -0.019	’10’ -0.004	’Hudson’ -0.003
’Ohio ’ 0.002	’280’ -0.019	’59’ 0.004	’98’ 0.003
’Farm’ -0.002	’1226’ 0.019	’JANET’ -0.004	’Cedar’ 0.003
’House’ -0.002	’Langford’ -0.019	’Arrow’ -0.004	’37’ -0.003
’MACAULEY’ -0.002	’6’ 0.019	’48’ 0.004	’Old’ -0.003
’45’ 0.002	’2505’ 0.019	’50’ 0.004	’10’ 0.003
’Tree’ -0.002	’24’ -0.019	’Soundview’ 0.003	’Tree’ -0.003
’School’ 0.002	’75’ -0.019	’98’ 0.003	’82’ -0.003
’24’ 0.002	’Chestnut’ 0.018	’W ’ -0.003	’7’ -0.003
’1226’ 0.002	’1294’ 0.018	’750’ 0.003	’W ’ -0.003
’Canterbury’ 0.002	’Sperry’ -0.017	’Cutlery’ -0.003	’High’ -0.003
’404’ -0.002	’WALNUT’ 0.017	’2768 ’ 0.003	’E’ -0.003
’Whitney’ 0.002	’Ridge’ 0.017	’Dr ’ 0.003	’Gertrude’ 0.003
’Sharon’ 0.002	’37’ 0.017	’555’ -0.003	’Pape’ -0.002
’ST’ -0.002	’Weed’ -0.017	’94’ -0.003	’Portland’ -0.002
’WILSON ’ -0.002	’Ln ’ -0.016	’JARVIS’ 0.003	’Oates’ -0.002
’Ave ’ 0.002	’ALYCE’ -0.016	’Meredith’ -0.003	’472’ -0.002
’28’ -0.002	’TPKE’ -0.015	’Mayflower’ -0.003	’Sharon’ -0.002
’Ln ’ 0.002	’Farmington’ -0.015	’Cedar’ -0.003	’238’ 0.002
’Sycamore’ 0.002	’28’ -0.015	’Myrtle’ -0.003	’750’ -0.002
’Ford’ -0.002	’Pendleton’ 0.015	’Beach’ 0.003	’Hartford’ -0.002
’317’ -0.002	’Sunny’ -0.014	’Ct ’ -0.003	’1’ -0.002
’12’ 0.002	’Mountain’ -0.013	’29’ -0.003	’Meredith’ -0.002

Table 20: LIME Results - Companies (1)

Baseline NN	LSTM	w/ Dates	w/ Letters
LIMITED -0.667	LIMITED -0.039	LIMITED -0.260	LIMITED -0.336
BOWKER 0.098	PROPERTIES -0.031	W 0.074	LTD -0.097
W 0.097	LIMITED.I.C. -0.021	A 0.048	PROPERTIES -0.070
A 0.070	BLUEBERRY 0.019	PROPERTIES -0.047	W 0.059
LIMITED.I.C. 0.069	PROPERTY -0.017	BOWKER 0.040	A 0.055
CROWTON 0.056	CROWTON 0.016	INNOVATIONS 0.038	BLUEBERRY 0.049
VISION 0.046	LTD -0.015	EMPLOYMENT -0.029	BOWKER 0.038
BLUEBERRY 0.041	EMPLOYMENT -0.011	ROAD -0.026	LIMITED.I.C. 0.036
SERVICES -0.033	SERVICES -0.010	TRANS 0.024	CROWTON 0.028
GROUP -0.033	BODZIO 0.010	COCHRANE -0.023	RATED 0.026
CHESHIRE -0.032	HEADQUARTERS 0.009	ACER 0.020	ARTSPACESLTD 0.025
EMPLOYMENT -0.030	SOLUTIONS -0.009	BSEC 0.020	SOLUTIONS 0.023
GOURMET 0.028	ORTHODONTICS 0.008	BODZIO -0.019	GLASGOW -0.023
DE-TOY 0.026	GOURMET 0.008	ENGINEERS -0.019	PROPERTY -0.022
RECLAIM 0.025	W 0.007	ATOS 0.019	HENRY 0.022
CHARLIE -0.025	BLINK -0.006	ANTHONY -0.019	BODZIO -0.020
DARWIN 0.022	MANPOWER 0.006	PROPERTY -0.019	RECLAIM -0.020
CONTRACTOR 0.022	M -0.006	ANDREI -0.018	COMPUTERS 0.019
PROPERTY -0.021	FALCON 0.006	BLUEBERRY 0.018	GAS -0.019
DR 0.021	LIGHTINTHEBOX -0.006	MINISTRY -0.017	ALAMORT 0.019
WINGS -0.021	DADDIES 0.006	COST -0.017	FUN 0.018
BENICKY 0.021	CYSEC 0.006	GOURMET 0.017	VEHICLE -0.018
GREEN 0.020	CLIVE 0.006	JOINT 0.016	CONSTANCE 0.018
AZAR 0.020	CRADDOCK 0.005	LTD -0.016	BLINK 0.018
ATOS 0.020	RATED 0.005	CAPITAL -0.016	DOBINSON -0.018
ATOMIC 0.020	RECLAIM 0.005	MANAGEMENT 0.016	MANPOWER -0.018
EQUITY -0.020	ALAMORT -0.005	PUBLISHING -0.015	LIMI 0.017
ADRIAN 0.019	BADER -0.005	IMAGING 0.015	M -0.017
PROPERTIES -0.018	BSEC -0.005	78 0.014	CRADDOCK 0.017
ANNE -0.018	BURUNGA 0.005	BELTA 0.014	PLUMBING -0.017
COLRON 0.018	CARDPRIZE -0.005	PROCESSING -0.014	78 -0.017
AJC -0.018	LEWISHAM 0.005	MUSIC -0.014	GATE -0.017
KITE -0.018	AZAR -0.005	MIDLANDS 0.014	PRODUCTIONS 0.016
LTD 0.017	CREATIVE 0.005	DR -0.013	DO -0.016
DAMICO -0.017	ARTSPACESLTD -0.005	OUTSKIRTS 0.013	EVENTS 0.016
CRADDOCK -0.017	VISION -0.005	EQUITY -0.013	PUBLISHING 0.016
BADER -0.017	TWO 0.005	BRIAN 0.013	CONTRACTOR 0.016
ACADEMY 0.017	CONTRACTOR 0.005	TWO -0.012	CELERUM -0.016
LIMITEMITED -0.017	BABBACOMBE 0.005	INFORMATION 0.012	ELITE 0.015
INNOVATIONS 0.017	78 -0.005	1 0.012	CELLWIZE -0.015
HOLDINGS 0.017	DOBINSON -0.005	BUSINESS 0.012	INNOVATIONS -0.015
CUM -0.017	COCHRANE -0.005	GIFTS 0.011	TRANSPORT -0.015
LIM -0.016	M. 0.005	BADER 0.011	PARTNERS 0.015
BIN 0.016	MANAGEMENT -0.005	HENRY -0.011	COCHRANE -0.014
VEHICLE 0.016	CELTIC 0.004	BOOKMAKERS 0.011	DBA 0.014
MUSIC -0.015	BUSINESS 0.004	CAR 0.011	A2B -0.014
AGCAS 0.015	GIFTS -0.004	FALCON 0.011	ELECTRICAL 0.014
SITE -0.015	JOE 0.004	ELECTRICAL 0.011	GOURMET 0.013
RESTAURANT -0.015	ALAIN 0.004	COMPUTERS -0.011	TRANS 0.013
CARLIN -0.015	BUCKLEFIELDS 0.004	CARLIN -0.011	BENICKY 0.013

Table 21: LIME Results - Companies (2)

Baseline NN	LSTM	w/ Dates	w/ Letters
ARTSPACESLTD -0.015	and 0.004	CHARLES 0.011	RESTAURANT -0.013
GAS 0.015	COST -0.004	COVENTRYTED 0.011	ROAD -0.013
M. 0.015	EQUITY -0.004	COLBERRY -0.010	MAN -0.013
CELLWIZE -0.015	CHARLES -0.004	AHL 0.010	ACADEMY 0.012
WORKS 0.014	APJ -0.004	BREW 0.010	LEWISHAM 0.012
A.V. 0.014	COMMUNITY 0.004	M. -0.010	DCandV 0.012
TRANSPORT 0.014	TRANS 0.004	RECLAIM -0.010	CAROLINE -0.012
CAR -0.014	ANTHONY 0.004	BOYLE -0.010	GOLF -0.012
HARDY 0.014	CDCE -0.004	LEWISHAM -0.010	BIN -0.012
COCHRANE -0.014	K 0.004	AJC 0.010	BIECO -0.012
AZK -0.014	VERONICA 0.004	M 0.010	EMPLOYMENT -0.012
PRODUCTIONS -0.013	UK -0.004	CELERUM -0.010	CONTRACTING -0.011
COST -0.013	LTD. -0.004	CANTINHO 0.009	ALYASAMEEN -0.011
GATE 0.013	PLUMBING -0.004	CROSSFIELDS -0.009	LIVING -0.011
BUTTERFLY -0.013	LIVING -0.004	MITSUBISHI -0.009	CANTINHO -0.011
FALCON -0.013	ACER -0.004	34 0.009	GOIAS 0.011
GLASGOW -0.013	AGCAS 0.004	CROWTON 0.009	BRIAN 0.010
TWO 0.013	AP -0.004	BRUNSWICK -0.009	ASSIST 0.010
ELTON 0.012	NATIONAL -0.004	AZK 0.009	TRAVEL 0.010
34 0.012	CONTRACTING -0.004	RUTLAND -0.009	ENGINEERS -0.010
BOOKMAKERS -0.012	AHL 0.004	APJ -0.009	ANTHONY 0.010
ALEEPH 0.012	VEHICLE -0.004	ANNE 0.009	WINGS 0.010
CYSEC -0.012	CROSSFIELDS -0.004	COMPANY 0.008	ORTHODONTICS 0.010
DEVELOPING -0.012	DARWIN 0.004	HOLDING -0.008	COST 0.010
CONISTON -0.012	BUTTERFLY 0.004	DARWIN -0.008	LIMITEDMPANY 0.010
ENGINEERING 0.012	ANDREI -0.004	STORES 0.008	CORNER 0.010
DEVELOPMENTS -0.012	BRUNSWICK 0.004	ASSIST 0.008	LIMITEMITED 0.010
LTD -0.012	INC -0.004	CHESHIRE -0.008	A.V. 0.010
AUTOMOTIVE 0.011	DESIGN -0.004	CEILINGS 0.008	ADVANCE 0.009
DBA 0.011	ADRIAN -0.003	164 0.008	ENGINEERING -0.009
ALAIN -0.011	JOINT -0.003	DE-TOY -0.008	AHL -0.009
BURGERS -0.011	COMPUTERS -0.003	CHORLTON -0.008	CONTEMPORARY -0.009
COMMUNITY 0.011	BURGERS 0.003	WINGS 0.008	P. 0.009
LEWISHAM -0.011	SPICE 0.003	2017 0.008	CELTIC -0.009
AND 0.011	MITSUBISHI -0.003	PLUMBING 0.008	AGE -0.008
PUBLISHING -0.011	WINGS -0.003	A. 0.008	ANNE 0.008
INTERIORS -0.011	CHEAM -0.003	CHARIOT -0.008	PRINT 0.008
and 0.011	HOLDINGS 0.003	ENGINEERING 0.007	CREATIVE -0.008
LLP -0.011	DEVELOPMENTS -0.003	PARTNERS -0.007	INTERES 0.008
AP 0.010	LTD -0.003	OPTICIANS 0.007	34 -0.008
C 0.010	PLUMBING -0.003	SOLUTIONS -0.007	1 -0.008
M -0.010	PRODUCTIONS 0.003	LTD. 0.007	INFORMATION -0.008
JEWELLERS -0.010	CRAGFIT 0.003	VENTURES 0.007	ACTU 0.008
VIATOR 0.009	LIMITED6LIMITED -0.003	HEATING -0.007	THE 0.007
CO -0.009	DIAMOND 0.003	ACTU -0.007	AZK -0.007
ALYASAMEEN 0.009	CARLIN 0.003	NATIONAL 0.007	IMAGING 0.007
BODZIO 0.009	CAROLINE -0.003	12 -0.007	PIMPEC -0.006
IMAGING 0.008	DEVELOPING -0.003	DEVELOPING -0.007	ALEEPH 0.006
CCI 0.008	PROCESSING -0.002	GAS 0.006	ASCURLO -0.006
ANDREI 0.008	EVENTS -0.002	COFFEE 0.006	MIDLANDS -0.006

Table 22: LIME Results - Names (1)

Baseline NN	LSTM	w/ Dates	w/ Letters
SAN 0.003	VIRELLA -0.116	SAN 4.42E-05	Theresa 0.0004
JOSE 0.003	COFFEY -0.115	Barker 4.36E-05	S 0.0003
DELA 0.003	D -0.107	W -4.28E-05	A 0.0003
BUCHANAN 0.003	K -0.101	COFFEY -3.73E-05	JOSE 0.0003
BONNER 0.003	M -0.086	Matthew 3.64E-05	Glen 0.0003
HICKMAN -0.003	L -0.085	Stephanie -3.48E-05	K -0.0003
MERCADO 0.003	JOSE -0.085	BARUCCHERI -3.43E-05	VIRELLA 0.0003
N -0.003	N -0.074	MATHIEU -3.41E-05	Valenzuela -0.0003
Alvarez 0.002	SAN -0.069	J 3.39E-05	Cain -0.0003
Gregory -0.002	MERCADO 0.044	Cain -3.38E-05	SAN 0.0003
C -0.002	SHAH 0.034	Alvarez 3.27E-05	Marlene -0.0003
REYNOLDS 0.002	S -0.032	Donaldson 3.25E-05	MERCADO 0.0003
BARNABY 0.002	Valenzuela 0.032	Payne 3.11E-05	Phillip -0.0003
Juanita 0.002	PEREZ 0.031	Darren -3.10E-05	Matthew -0.0002
VIRELLA 0.002	ANDERSON -0.028	Derek 3.09E-05	L -0.0002
FAJARDO -0.002	Phillip -0.028	GRULLON 3.06E-05	DAYS -0.0002
GOMEZDELATORRE 0.002	Juanita 0.027	Herrera 2.95E-05	ANDERSON -0.0002
Tate -0.002	SCHILLING 0.025	Kathleen 2.95E-05	KUZMA 0.0002
E -0.002	Anita 0.025	JOSE 2.87E-05	Nathan 0.0002
IBERN 0.002	Sandoval -0.024	Kelly 2.83E-05	GRULLON 0.0002
Matthew -0.002	Kathryn -0.024	RESTREPO -2.83E-05	Kaufman -0.0002
SHAH -0.002	DELA -0.024	EDWARDS 2.82E-05	LEAKE -0.0002
COLLINS -0.002	Marlene 0.023	Hendrix 2.77E-05	Stephanie 0.0002
SADIQ -0.002	TOBIN -0.022	SANTIAGO -2.77E-05	M -0.0002
P 0.002	A 0.022	Tate -2.76E-05	MYRIE -0.0002
A 0.002	SHARELL 0.022	SHARELL 2.71E-05	B 0.0002
KUZMA -0.002	Shannon -0.021	A 2.70E-05	REYNOLDS -0.0002
Heath 0.002	Herrera 0.021	PEREZ -2.70E-05	DIZDAREVIC 0.0002
Kathleen -0.002	SANTIAGO 0.019	SANTORO 2.69E-05	EDKINS -0.0002
Virginia -0.002	Riggs -0.019	Sandoval 2.63E-05	J 0.0002
Theresa 0.002	Hurley -0.019	Megan 2.60E-05	YE -0.0002
MAGRAS 0.002	J -0.019	SADIQ 2.59E-05	Hannah 0.0002
Payne -0.002	JIMENEZ -0.019	DRAYCOTT 2.56E-05	Mark 0.0002
Darlene 0.002	Guzman 0.018	YE 2.55E-05	Paul -0.0002
Perkins -0.002	BARNABY -0.018	Heath 2.46E-05	Tommy 0.0002
Cain -0.002	Nathan -0.018	Morton 2.45E-05	SHARELL 0.0002
Mariah 0.002	Stephanie -0.017	Lott 2.44E-05	Alice -0.0002
SANTIAGO -0.001	REYNOLDS 0.017	Perkins 2.40E-05	Guzman 0.0002
Duffy -0.001	HARDIAL 0.016	OLIVO 2.36E-05	Chester -0.0002
BUTASEK 0.001	Allison 0.016	DELA 2.36E-05	Walters -0.0002
Allen -0.001	LIMATO 0.016	DEVITO-RODRIGUE 2.33E-05	Brooke -0.0002
RESTREPO 0.001	SANTANGELO -0.016	BUTASEK 2.31E-05	Perkins 0.0002
Galloway 0.001	YE 0.016	H 2.27E-05	Bass -0.0002
Moss -0.001	GRULLON 0.016	R 2.23E-05	BARUCCHERI -0.0002
PEDROSA 0.001	Z 0.015	N 2.20E-05	MORENO -0.0002
Tessa 0.001	Morton 0.015	Brooke 2.20E-05	SANTANGELO 0.0002
DRAYCOTT 0.001	GARCIA -0.015	Stevenson 2.18E-05	V -0.0002
Anita -0.001	Teresa -0.014	D -2.15E-05	WILLIAMS 0.0002
Phillip -0.001	EDWARDS 0.014	C 2.13E-05	Garrett 0.0002
Love -0.001	Gregory 0.014	GARCIA 2.12E-05	Donaldson 0.0002

Table 23: LIME Results - Names (2)

Baseline NN	LSTM	w/ Dates	w/ Letters
Valenzuela 0.001	WLADIS -0.014	Hurley 2.10E-05	GRULLON 0.0002
COFFEY 0.001	Bass 0.014	ISLA 2.10E-05	IBERN 0.0002
Barker -0.001	Payne 0.014	Jack 2.08E-05	Roger 0.0002
Melendez 0.001	Love -0.014	SCHILLING -2.07E-05	SANTORO -0.0002
B -0.001	H -0.014	Elizabeth 2.07E-05	Kelly 0.0002
TAVERAS -0.001	Elizabeth 0.014	DIZDAREVIC 2.05E-05	Hatfield 0.0002
Herrera -0.001	C -0.014	P 2.04E-05	Jack 0.0002
EDKINS -0.001	GOMEZDELATORRE -0.013	Darlene -2.04E-05	Anita 0.0002
Edwards 0.001	MATHIEU -0.013	MERCADO 2.02E-05	Sandoval 0.0002
R -0.001	REID -0.013	COLLINS 2.02E-05	JIMENEZ 0.0002
REID 0.001	Arthur -0.013	Allison 1.98E-05	EDWARDS -0.0001
Alice 0.001	Glen 0.013	LARRY -1.97E-05	Alvarez 0.0001
LAZARO 0.001	Shannon -0.013	VIRELLA 1.96E-05	PEDROSA -0.0001
Paul 0.001	Brandi -0.013	Brandi 1.96E-05	Lott 0.0001
Y 0.001	Paul -0.013	SHAH 1.88E-05	Shannon 0.0001
Brooke -0.001	B -0.013	Brady 1.86E-05	Gay 0.0001
GARCIA 0.001	Derek -0.013	Duffy 1.86E-05	Sears 0.0001
LIMATO -0.001	DEVITO-RODRIGUE 0.012	IBERN 1.85E-05	Chase -0.0001
BARUCCHERI -0.001	Perkins -0.012	BONNER 1.82E-05	TOBIN -0.0001
Arnold -0.001	LEAKE -0.012	Melendez -1.80E-05	Hendrix 0.0001
MYRIE 0.001	Alvarez 0.012	SANTANGELO -1.73E-05	MATHIEU 0.0001
Arthur -0.001	BRESNAHAN 0.011	K 1.69E-05	Lester 0.0001
ISLA 0.001	Mariah 0.011	M -1.66E-05	Allen 0.0001
Mark -0.001	Matthew 0.011	GRULLON -1.66E-05	H 0.0001
V 0.001	Lester -0.011	Paul 1.62E-05	Darlene 0.0001
Gay 0.001	KUZMA 0.011	HERSHBERGER 1.59E-05	HARDIAL -0.0001
Danny 0.001	Donaldson -0.010	TOBIN 1.59E-05	SCHILLING -0.0001
YE -0.001	Barker 0.010	Tommy 1.58E-05	Carolyn -0.0001
WILLIAMS -0.001	R -0.010	Whitley -1.58E-05	Tate 0.0001
Hannah 0.001	IBERN -0.010	TAVERAS -1.53E-05	REID 0.0001
F 0.001	LAZARO 0.010	Shannon 1.52E-05	Clayton 0.0001
Cooke -0.001	TAVERAS 0.010	KUZMA 1.52E-05	ISLA 0.0001
Hatfield -0.001	MYRIE -0.010	BARGLOWSKA 1.51E-05	Megan -0.0001
Teresa -0.001	Heath -0.009	Edwards -1.51E-05	SOTO -0.0001
LARRY 0.001	FERONE -0.009	Valenzuela 1.50E-05	Whitley -0.0001
Chester 0.001	F 0.009	Virginia -1.46E-05	BARNABY -0.0001
Stephanie 0.001	Tessa -0.008	SOTO 1.44E-05	GOMEZ 0.0001
Kelly -0.001	Gay 0.008	Kathryn 1.43E-05	Elizabeth -0.0001
Carolyn -0.001	Galloway 0.008	Glen 1.43E-05	T -0.0001
DEVITO-RODRIGUE -0.001	Jack -0.008	MORENO 1.42E-05	COFFEY -0.0001
BARGLOWSKA -0.001	DIZDAREVIC 0.008	Z -1.39E-05	HICKMAN -0.0001
T 0.001	E 0.008	Gregory 1.37E-05	HERSHBERGER 0.0001
Brandi -0.001	G -0.008	Allen 1.34E-05	Barker -0.0001
Guzman -0.001	BUTASEK -0.007	MYRIE 1.33E-05	Shannon -0.0001
Chase -0.001	MAGRAS 0.007	GOMEZ -1.33E-05	BUCHANAN -0.0001
Brown -0.001	WILLIAMS 0.007	CAMPBELL 1.25E-05	C 0.0001
FERONE -0.001	Allen -0.007	WILLIAMS -1.24E-05	BUTASEK 0.0001
Z -0.001	MORENO -0.007	Y -1.19E-05	Morton -0.0001
Allison -0.001	Roger 0.006	Mark -1.18E-05	Moss -0.0001
Roger 0.000	BUCHANAN -0.006	S 1.06E-05	Brady -0.0001

Table 24: LIME Results - Products (1)

Baseline NN	LSTM	w/ Dates	w/ Letters
Hips -2.86E-06	Blender 7.37E-06	fluid 1.82E-05	60 -2.76E-06
32 -2.79E-06	Bar -4.36E-06	Surimi 1.73E-05	Traditional 2.66E-06
Support -2.71E-06	Moisturizing -2.10E-06	Protect-A-Bed 1.65E-05	Free -2.37E-06
Cheese -2.56E-06	Waring -2.04E-06	Mixed -1.62E-05	Travel 2.32E-06
3 -2.56E-06	Mushroom 1.91E-06	Greek -1.52E-05	Moisturizing -2.01E-06
lb -2.51E-06	Hips 1.36E-06	L-lysine 1.51E-05	Skin -1.93E-06
pack,1 -2.45E-06	tabs -1.11E-06	5pcs 1.51E-05	Mite -1.88E-06
3/ 2.44E-06	Lime -1.06E-06	Blender -1.47E-05	Carmel 1.87E-06
Eggo -2.43E-06	Lite 1.02E-06	Blackforest -1.43E-05	Fleurage 1.80E-06
Cranberry -2.43E-06	Fo 9.68E-07	Candle -1.43E-05	Blue -1.77E-06
Food -2.43E-06	chair -8.35E-07	Rhino 1.43E-05	Intensite 1.72E-06
Tubes -2.42E-06	With -8.19E-07	GRS.| -1.42E-05	Hair -1.70E-06
Reducer -2.42E-06	1000 -7.70E-07	system 1.42E-05	German -1.68E-06
Seed -2.39E-06	Foxtail 7.67E-07	”Vitamin 1.41E-05	Berry -1.68E-06
vcaps -2.38E-06	Toothbrush -7.22E-07	Biphosphate 1.40E-05	chair -1.64E-06
6 -2.36E-06	vcaps -7.20E-07	Reducer 1.40E-05	98% -1.63E-06
BACON -2.33E-06	Protector -7.20E-07	Peanut 1.38E-05	Mushroom -1.63E-06
Style -2.32E-06	Purple -7.02E-07	Waterlily 1.38E-05	10 -1.62E-06
Box -2.30E-06	Tea -6.92E-07	Supremo 1.38E-05	Facial -1.59E-06
98De -2.26E-06	STYLE -6.81E-07	10 -1.36E-05	Homocysteine -1.57E-06
Liquid -2.25E-06	Vermont -6.76E-07	MELHORAL 1.36E-05	P 1.56E-06
5 -2.24E-06	Exceptional 6.75E-07	Tv 1.36E-05	Makeup -1.51E-06
Tablets -2.22E-06	- -6.73E-07	130 -1.35E-05	Tabs -1.50E-06
candle -2.21E-06	Awakening -6.68E-07	Tracker -1.34E-05	Packet -1.50E-06
oz -2.20E-06	crabe -6.67E-07	Amish -1.33E-05	6x100 -1.48E-06
Soft -2.18E-06	Fever -6.63E-07	piece 1.33E-05	Dry -1.47E-06
Creme -2.14E-06	Daily 6.60E-07	Wafers 1.32E-05	Protein -1.47E-06
Ages -2.11E-06	Berry -6.57E-07	With -1.31E-05	Greek -1.46E-06
Cooking -2.10E-06	Height -6.56E-07	Dog -1.29E-05	Rtu -1.46E-06
Facial -2.08E-06	Metal -6.55E-07	Moxie -1.29E-05	Cocktail -1.45E-06
2 -2.08E-06	Peach -6.47E-07	Carmel -1.29E-05	LECHE -1.44E-06
25/4UNS -2.07E-06	Of -6.41E-07	Meatloaf 1.27E-05	Bit -1.44E-06
With -2.06E-06	Madness 6.39E-07	Bariatric 1.27E-05	182 -1.41E-06
Plus -2.06E-06	ALOE -6.37E-07	Puzzle 1.27E-05	Disney -1.41E-06
Seltzer -2.03E-06	Peanut -6.29E-07	Action 1.26E-05	C -1.40E-06
23123 -1.99E-06	Mighty -6.24E-07	Colombian 1.26E-05	Egg -1.37E-06
Core -1.95E-06	Egg -6.23E-07	ESPUMA 1.26E-05	kg -1.36E-06
Conditioner -1.94E-06	mg, 250 -6.22E-07	Macaroni 1.25E-05	Achari -1.35E-06
Hair -1.93E-06	Even -6.18E-07	Conditioner -1.25E-05	Spaetzle -1.35E-06
Flute -1.87E-06	creme -6.18E-07	Lite -1.25E-05	Waffles -1.35E-06
Iced 1.86E-06	Meatloaf -6.09E-07	6 -1.25E-05	Sinus -1.35E-06
150 1.83E-06	Enamel -6.07E-07	Pain -1.24E-05	Teva 1.33E-06
Duck -1.69E-06	Traditional -5.95E-07	Noodle -1.23E-05	Meatloaf -1.32E-06
chair 1.52E-06	Just -5.77E-07	Double-sided -1.22E-05	Mattress -1.29E-06
Tonic 1.46E-06	70% -5.76E-07	Origins -1.21E-05	BABARIA|BABARIA -1.27E-06
Teva 1.38E-06	Margarita -5.72E-07	Gourmet 1.21E-05	Dress -1.25E-06
Height 1.36E-06	Splash -5.66E-07	Set -1.21E-05	piece -1.25E-06
FRUCTIS|FRUCTIS 1.33E-06	Tubes -5.60E-07	- -1.20E-05	Primer -1.24E-06
1x100 1.31E-06	Honey -5.44E-07	X -1.20E-05	Tea -1.24E-06

Table 25: LIME Results - Products (2)

Baseline NN	LSTM	w/ Dates	w/ Letters
each 1.26E-06	3.6 -5.43E-07	6x100 -1.20E-05	Fat -1.23E-06
Ham 1.25E-06	Cooking -5.32E-07	Homocysteine -1.20E-05	crabe 1.21E-06
Foot 1.23E-06	Fruit -5.32E-07	Predator -1.19E-05	Caplets -1.21E-06
Game 1.21E-06	De -5.26E-07	Fluffy -1.19E-05	One -1.20E-06
Football 1.20E-06	Drink -5.25E-07	creme -1.19E-05	Contemporary -1.19E-06
Bit 1.20E-06	SUN -5.22E-07	caplets -1.19E-05	Style -1.18E-06
Xylitol 1.11E-06	system -5.06E-07	Daily 1.18E-05	24 -1.18E-06
Sinus 1.11E-06	120 -5.05E-07	SABOR 1.18E-05	Height -1.18E-06
BRINOX 1.06E-06	Detoxitech -5.04E-07	Attractant -1.17E-05	ALOE -1.17E-06
Pain 9.97E-07	Sl400u 5.03E-07	Seed -1.17E-05	Focus -1.16E-06
Candle 9.94E-07	Core -4.97E-07	Eyeshadow -1.16E-05	candle -1.15E-06
capsule 9.90E-07	Miettes -4.90E-07	Facial -1.15E-05	Definition -1.15E-06
P 9.64E-07	Cocktail 4.90E-07	toothbrush -1.15E-05	Moxie -1.14E-06
LAMP 9.42E-07	4 4.85E-07	Dry -1.14E-05	American -1.14E-06
Purple 9.37E-07	Omega -4.83E-07	Bar -1.14E-05	Country -1.12E-06
Powder 9.15E-07	Action -4.83E-07	3.6 -1.13E-05	Waring -1.11E-06
Snow 8.88E-07	Oolong -4.72E-07	Tablets 1.13E-05	L -1.09E-06
Colour 8.78E-07	Almonds -4.69E-07	Focus -1.13E-05	Holiday -1.07E-06
SABOR 8.31E-07	500 -4.69E-07	Colour 1.13E-05	Normal -1.05E-06
Sodium 7.82E-07	Ages 4.68E-07	Soft -1.12E-05	Seltzer -1.04E-06
Elbow 7.81E-07	Spray 4.64E-07	Juice -1.11E-05	China 1.03E-06
Up 7.67E-07	de -4.63E-07	count -1.09E-05	Madness -1.02E-06
Counter 7.49E-07	Blackforest 4.55E-07	BWF -1.07E-05	tabs -9.90E-07
ALOE 7.29E-07	Chewing -4.46E-07	Vermont 1.05E-05	toothbrush -9.71E-07
Colombian 6.99E-07	Gourmet 4.45E-07	savon -1.05E-05	Just -9.43E-07
Country 6.59E-07	Gourmet 4.23E-07	Oregano 1.04E-05	Oolong -8.79E-07
Attractant 6.21E-07	Panel -4.12E-07	Kosher 1.04E-05	Anti-bacterial -8.66E-07
Skin 6.02E-07	Lung 4.12E-07	Traditional 1.03E-05	Foxtail 8.51E-07
F.50+AFTER|+AFTER 5.98E-07	Formula -4.02E-07	0.2-0.5% 1.03E-05	Glass 8.49E-07
Metal 5.70E-07	182 4.00E-07	To 1.03E-05	Size 8.12E-07
Drops 5.55E-07	savon 3.98E-07	Bristles 1.02E-05	Root -7.91E-07
Gum 5.55E-07	Unit -3.92E-07	Seasoned -1.02E-05	Drops 7.80E-07
Blackforest 4.95E-07	Caplets 3.89E-07	Ecofam -1.01E-05	Noodle 7.53E-07
Rtu 4.88E-07	in -3.83E-07	P -1.01E-05	Shape 7.51E-07
Soak 4.84E-07	33 3.69E-07	Sl400u -1.01E-05	90 6.93E-07
Intensite 4.17E-07	Elbow 3.65E-07	Drops -9.77E-06	Pain 6.92E-07
1/2-inch 4.07E-07	Smoothing 3.37E-07	Football 9.55E-06	Ham 6.91E-07
Tablets 4.06E-07	Seltzer 2.90E-07	gr -9.49E-06	Up 6.36E-07
PHILIPS 3.98E-07	Back 2.83E-07	Game -9.49E-06	gevrey 6.34E-07
Factors 3.78E-07	COLHER 2.79E-07	500 -9.27E-06	each 6.27E-07
Mfg. 3.65E-07	Allergy 2.65E-07	Velour -9.25E-06	Tonic 6.15E-07
X 2.88E-07	vegetarian 2.62E-07	Of 9.13E-06	Clean 5.76E-07
Formula 2.64E-07	Enema 2.54E-07	tabs 8.76E-06	Lite 5.50E-07
LECHE 2.56E-07	Facial 2.51E-07	FRUCTIS|FRUCTIS -8.71E-06	Series 5.46E-07
Lemonade 2.54E-07	6 2.29E-07	Flute 8.68E-06	Feet 5.39E-07
American -2.39E-07	In 2.00E-07	150 8.44E-06	FRUCTIS|FRUCTIS 5.32E-07
Gotcha 1.87E-07	Reducer 2.00E-07	Makeup 8.16E-06	savon 5.22E-07
Surimi 1.74E-07	candle 1.91E-07	Tomato -8.06E-06	Large 5.16E-07
Ecofam 6.56E-08	chamb.dom.trapet 1.82E-07	mg,250 7.82E-06	Gel 3.93E-07
Focus 5.50E-08	Rtu 1.20E-07	Counter 7.79E-06	Rose 3.58E-07

Table 26: Precision/Recall- Addresses

Embeddings	precision	recall	f1-score	# instances tested
GloVe,d=50	.95	.99	.97	3586
GloVe, d=100	.94	.95	.95	3586
GloVe, d=200	.94	.99	.96	3586
GloVe, d=300	.95	.99	.97	3586
GloVe 42B, d=300	.95	.99	.97	3586
CBOW(n=2)	.95	.99	.97	3586
CBOW(n=5)	.95	.99	.97	3586
Dependancy Based	.96	.99	.97	3586
fastText	.93	.99	.96	3586
Character Embedding	.96	.99	.98	3586
Character Softmax Embedding	.95	.99	.97	3586

Table 27: Precision/Recall- Companies

Embeddings	precision	recall	f1-score	# instances tested
GloVe,d=50	.71	.78	.74	4019
GloVe,d=100	.70	.78	.74	4019
GloVe,d=200	.69	.77	.73	4019
GloVe,d=300	.68	.78	.73	4019
GloVe 42B,d=300	.82	.51	.63	4019
CBOW(n=2)	.71	.78	.74	4019
CBOW(n=5)	.70	.79	.74	4019
Dependancy Based	.70	.80	.74	4019
fastText	.81	.51	.63	4019
Character Embeddings	.72	.46	.56	4019
Character Softmax Embedding	.73	.46	.57	4019

Table 28: Precision/Recall- People

Embeddings	precision	recall	f1-score	# instances tested
GloVe,d=50	.89	.80	.84	5859
GloVe,d=100	.89	.79	.74	5859
GloVe,d=200	.88	.77	.73	5859
GloVe,d=300	.88	.76	.82	5859
GloVe 42B,d=300	.76	.92	.83	5859
CBOW(n=2)	.89	.80	.84	5859
CBOW(n=5)	.89	.78	.83	5859
Dependancy Based	.89	.78	.84	5859
fastText	.76	.91	.83	5859
Charcter Embedding	.72	.89	.80	5859
Character Softmax Embedding	.72	.89	.80	5859

Table 29: Precision/Recall- Products

Embeddings	precision	recall	f1-score	# instances tested
GloVe,d=50	.87	.89	.88	2976
GloVe,d=100	.87	.89	.88	2976
GloVe,d=200	.87	.89	.88	2976
GloVe,d=300	.86	.89	.88	2976
GloVe 42B,d=300	.86	.90	.88	2976
CBOW(n=2)	.87	.89	.88	2976
CBOW(n=5)	.87	.90	.88	2976
Dependancy Based	.88	.89	.89	2976
fastText	.86	.88	.87	2976
Character Embedding	.89	.85	.87	2976
Character Softmax Embedding	.88	.87	.87	2976