Evaluating Food-Drug Interactions with Artificial Neural Networks
UCB W266 Summer 2017 Project Summary

Adam Lenart, Hyera Moon, Lisa Barceló

Abstract

The increasing number of Americans managing multiple medications to treat chronic disease necessitates the efficient exploration of food-drug interactions. This study used a variety of models (recursive neural network, bag-of-words logistic regression classifier, and long short term memory neural network) to identify key interactions between drug classes and common food compounds.

Keywords — NLP; Natural Language Processing; Information Extraction; Pharmacovigilance; Food-Drug Interactions; LSTM; Long Short-Term Memory

1 Introduction

An estimated 40 million Americans use prescription medication to control hypertension, diabetes, and a variety of comorbidities (Qato et al., 2008; Go et al., 2014). Of these, approximately 10 million are managing multiple chronic diseases (Gerteis et al., 2014). Medication management is a healthcare best practice, and patients and providers alike need to be educated on the risks of drug-drug interactions that can occur when managing medications for a variety of comorbidities. While the FDA has released guidelines on the most common known food-drug interactions[1], hundreds of new compounds are being tested and released each year and good practice in pharmacovigilance requires continuously updating the guidelines (FDA, 2005). More traditional approaches to pharmacovigilance encompass manual investigation of the literature (Holbrook et al., 2005; Bushra and Yar Khan, 2011). However, as the number of biomedical articles[2] increases by about 4.5% each year, leading to a doubling time of about every 15 years, this approach is not scalable.

Aiming at automated food-drug interaction recognition, we used neurolinguistic processing tools to shed light on relationships between specific drug classes and nutrients found in common food items, both naturally occurring (i.e., isoflavonoids and plant estrogens) or synthetic additions (i.e., monosodium glutamate and bisphenol-A). Advances in technology are bringing about an increased focus to the application of artificial neural networks (ANN) in biomedical research (Suzuki, 2011). These advances can enable researchers to solve healthcare's most pressing questions. Using the body of relevant research articles, data mining to identify potentially harmful interactions between a given drug class and food compounds can be enhanced through the use of ANN or other Natural Language Processing models.

2 Background

Biomedical natural language processing applications often focus on finding adverse events (Kang et al., 2014; Culbertson et al., 2014; Maitra et al., 2014), gene expression statistics (Karopka et al., 2004; Geifman et al., 2015; Chen et al., 2014) or drug-disease pairs (Xu and Wang, 2013; Wang et al., 2017a; Wang et al., 2017b). However, efforts to elucidate food

[1] Avoid Food-Drug Interactions: https://www.fda.gov/downloads/Drugs/.../GeneralUseofMedicine/UCM229033.pdf

[2] PubMed citation statistics: https://www.nlm.nih.gov/bsd/licensee/baselinestats.html

and drug interactions based on the medical literature have been scarce; only Zhang et al. (2015) attempted to uncover them. While the work of Zhang et al. (2015) made use of rule-based semantic predictions on the PubMed corpus, building on work by Rindflesch and Fiszman (2003), we used ANNs to forego the usage of rules.

3 Data

PubMed is a collection of 27 million[3] citations for the biomedical literature. As it also includes RESTful APIs which allow easy access to the metadata of these articles and FTP access to Open Access articles, PubMed is an often used resource for Natural Language Processing applications (Hunter and Cohen, 2006).

In this paper, we focused on 12 drug classes chosen to represent a wide spectrum of treatment areas (Table 3). PubMed included over 300,000 citations on 2017-AUG-21 corresponding to these drug classes (Fig. 1). We downloaded the corresponding abstracts and parsed them with the NLTK Python library, keeping only the sentences containing our key word (drug name). The PubMed search engine uses Automatic Term Mapping in its search, and includes multiple variations of compound names in its search; for example, a search for *levothyroxine* causes the search engine to also query the terms *laevothyroxine* and *thyroxine*. To account for this, we used fuzzy string matching and Jaro-Winkler distances to ensure we were capturing sentences with all relevant terms. We then filtered further by selecting only sentences where our key word co-occurred with a food compound. Our database of 20,000 common foods and the components came from FooDB.ca, which we pre-processed as a dictionary of the foods and their components, e.g., {milk: 'whey','casein'}. In total, we found 2,467 sentences containing both drug classes and 886 of the 20,000 possible food items.

Subsequent to extracting the sentences, we labeled them according to whether or not the interaction between the food compound and the drug in question was complementary. The resulting labels were either *positive, neutral* or *negative.* Table 3 in Appendix B shows the label distribution of the sentences. The average sentence length was 32 words, with the longest sentence containing 665 words.

4 Methods

We opted to implement a classification model in a supervised learning setting. The interaction between drugs classes and common food compounds are classified as one of three possible labels: *negative*, *neutral*, and *positive*.

Our baseline model used random prediction, assigning sentences one of three labels with equal probability. This method obtained an accuracy of 33.4%.

Following the recommendations of (Goodfellow et al., 2016, 413), as we had a limited number of labeled sentences available, we first used logistic regression and then moved on to a RNN-Long Short-Term Memory (LSTM) implementation. Additionally, we used another RNN model, the Stanford Sentiment Analysis, to estimate the direction of the interaction between food compounds and drugs.

4.1 Logistic Regression

We first began with logistic regression, a simple statistical model, before exploring the deep learning approach (Wang and Manning, 2012, 93).[4]

We initially hypothesized that the order of words in the sentences would not be important for detecting interaction and decided to use a bag-of-words model. More specifically, we used a bag-of-words formulation of logistic regression for classification using TF-IDF features. The frequency-based embedding using TF-IDF vectorization which would assign higher weights to words that appeared frequently, but would add a penalty if the word appeared in too many sentences, such as stop words.

In this model, we first constructed the word vectors, $\mathbf{W}$, according to

$$\mathbf{W} = \mathbf{tf} \log \frac{N}{\mathbf{idf}} ,$$

[3] Accessed on 2017-AUG-21.

[4] We also tried SVM in place of logistic regression but we found that logistic regression generalizes better to the validation set.

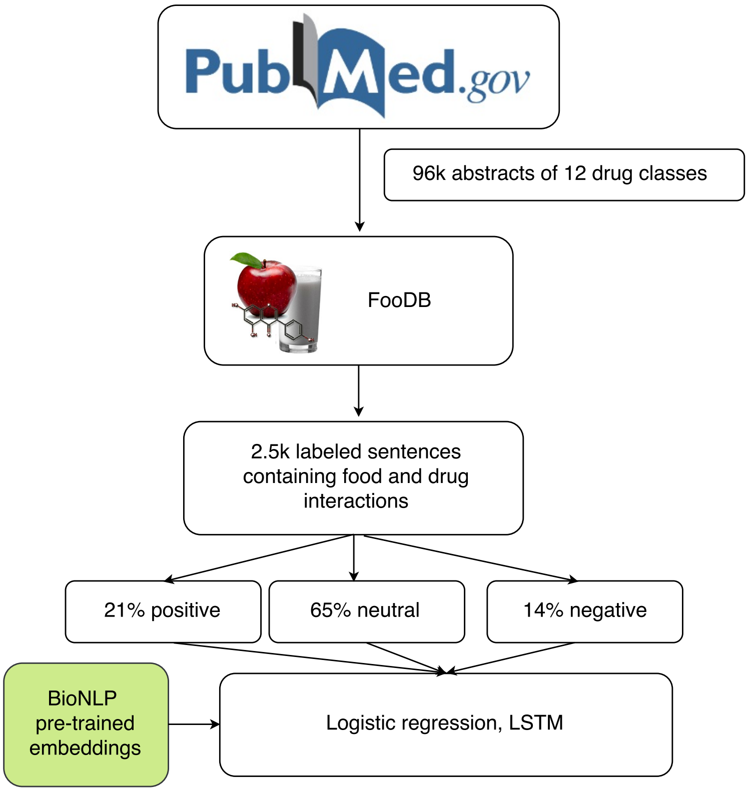

PubMed.gov
96k abstracts of 12 drug classes
FooDB
2.5k labeled sentences containing food and drug interactions
21% positive
65% neutral
14% negative
BioNLP pre-trained embeddings
Logistic regression, LSTM

Figure 1: **Project pipeline.** From more than 300,000 abstracts downloaded from PubMed, 2,467 sentences were extracted and labeled which contained food and drug interactions. Later, these sentences were processed by NLP algorithms.

where N denotes the number of words, **tf** the term-frequency vector for each word in each sentence, and *idf* the inverse document frequency as

$$\mathbf{idf}_i = \log \frac{|D|}{\{d : w_i \in d\}} ,$$

that is the inverse of the number of sentences where word w_i appears out of $|D|$ sentences. $\mathbf{W}$ can be used in estimating the conditional probability of $P(y_k|W)$ in the multilabel classification logistic regression directly as below:

$$P(y_k|\mathbf{W};\theta_k) = \frac{1}{1 + exp(-y_k\theta_k\mathbf{W})}$$

for label k.

Lastly, to avoid learning very large weights, which are likely to fit the training data but not generalize well, we used the L2 regularization.

4.2 RNN Sentiment Analysis Model

As Socher et al. (2013) describe, the RNN sentiment analysis model parses sentences into binary trees where each word of the sentence is represented in vector form as tree leaves, and the parent node vectors are computed recursively, bottom-up.

The word vectors are stored in a $d \times |V|$ embedding matrix where d stands for the embedding dimension and $|V|$ the vocabulary size. For word vector $\mathbf{v}$ the labels are given as

$$\mathbf{y} = softmax(\mathbf{Wv}) ,$$

where $\mathbf{y}$ is the output vector of five elements containing probabilities for the labels *very negative, somewhat negative, neutral, positive, very positive.* For the purposes of our study, we grouped *very negative* and *somewhat negative* labels as *negative*, and *very positive* and *somewhat positive* as *positive* labels, respectively.

In this model, we tested our initial hypothesis that a sentiment could be used as a proxy to detect interaction between a drug and food compound in a sentence. We used the Stanford Sentiment Analyzer's recursive neural network trained on a movie review corpus to evaluate the sentiments for each of our sentences.

4.3 Vanilla LSTM

LSTM was proposed by Hochreiter and Schmidhuber (1997) to efficiently solve long time-lag tasks. One cell consists of three gates (input, forget, output), and a cell unit. The LSTM cell can be defined by the following set of equations:

$$\begin{aligned}
input_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \\
forget_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \\
output_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \\
c_t &= forget_t \circ c_{t-1} + input_t \circ \\
&\quad tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\
h_t &= output_t \circ tanh(c_t) ,
\end{aligned}$$

where $input_t$, $forget_t$, $output_t$ denote the input, forget and output gates, respectively, c_t stands for the state of the cell and h_t stands for the output vector at word t. The LSTM output is channeled to three fully connected neurons denoting the probabilities for negative, neutral or positive predictions.

4.3.1 Hyperparameters

The vocabulary size for the Vanilla LSTM was 20,000 words and the additional food components that were not included in the top 20,000 words of the corpus, yielding a total of 20,554 words. The LSTM layer had the highest validation accuracy with 300 hidden units. The sentences were truncated or padded to 100 words.

4.4 2-layer LSTM

The 2-layer LSTM is similar to the Vanilla LSTM, the only difference is an additional second hidden layer. The motivation behind adding an additional LSTM layer is to allow the model to learn new semantic relationships that the single LSTM layer, being closer to the raw data, could not learn. The 2-layer LSTM model was augmented by dropout regularization (Srivastava et al., 2014; Goodfellow et al., 2016). To reduce the dimensionality of the problem, we added pre-trained BioNLP word embeddings[5] on the PubMed corpus to the model[6].

[5] BioNLP word2vec: http://bio.nlplab.org/

[6] We also trained skipgram word embeddings on Open Access articles of biopharm journals downloaded from the

Model	Accuracy (%)
Baseline - random	33
RNN - sentiment analysis	17
BOW LR_1	73
BOW LR_2	67
Vanilla LSTM	73
Vanilla LSTM pre-trained	61
2-layer LSTM	78
2-layer LSTM pre-trained	69

Table 1: **Model Accuracy**. BOW LR_1 denotes Bag-of-words logistic regression with split across all drug classes, BOW LR_2 denotes a bag-of-words logistic regression model trained on all but one drug classes, and tested on the drug class that was left out.

4.4.1 Hyperparameters

The best validation accuracy was reached with a vocabulary size of 30,000 words, as well as the additional food components not included in the top 20,000 words of the corpus, yielding a total of 30,499 words. The model had the highest accuracy when both LSTM layers had 300 hidden units and the dropout rate was 0.2.

5 Results

The models were trained on 75% of the sentences, 25% were conserved for testing. The accuracy of each model is shown in Table 1.

5.1 Evaluation

Due to the fact that we had unbalanced class distribution (65% neutral), accuracy was not a sufficient metric to evaluate our model efficacy. That is, a model that predicts 100% neutral sentences would still have an accuracy of 65%. Therefore, we evaluated our model not only with accuracy, but also with an F1 score, precision, and recall, using a confusion matrix to assess model performance with each label. Table 2 shows the percentage of incorrectly predicted labels with a higher percentage error for neutral label, except for the RNN sentiment analysis model which was not trained with the unbalanced labeled sentences.[7]

As Table 1 shows, the RNN Sentiment Analysis Model overwhelmingly predicted negative sentiment, and obtained an accuracy of 17.2%. Part of the disparity in the results relates to the vocabulary used for training. The frequency histogram for that model shows that as sentence length grows to 15 or more words, the sentiment distribution skews negative. Primarily, we observed that "sentiment" in the traditional sense is a poor judge of molecular interaction. For example, a food compound can cause upregulation of an inhibitor, working in concert with the medication (*positive* interaction), or competitively inhibit a known hormone agonist, working against the medication (*negative* interaction). These nuanced interactions are not likely to be reflected in a non-scientific training corpus.

Using a bag-of-words logistic regression classification, we were able to arrive at about 73% accuracy when the model encountered test sentences with drug class from the previously trained sentences. The accuracy decreased to 67% when the model encountered a new class of drugs it had not seen before.

5.2 Discussion

One of the challenges with our dataset is that we saw overlap between some common food names and non-food-related words in the scientific corpus. For example, searching for co-occurrences with the food "pie" initially included sentences containing the word "therapies". The Jaro-Winkler string matches were helpful in resolving this issue.

Another problem we encountered was incongruencies in word contexts. The food item "date" was indistinguishable from the "date" of an experiment, for example.

Our random predictor baseline has been recommended as a good comparison against which we can measure our resulting model. We will need to evaluate whether or not the model is better than chance at predicting the classification of the sentence.

PubMed FTP site but the BioNLP word2vec embeddings included more of our food items.

[7] For more detailed error analysis, refer to Table 5 and Figure 2

	Incorrectly Predicted As (%)		
Model	negative	neutral	positive
RNN - sentiment analysis	84	2	1
BOW LR_1	3	18	6
BOW LR_2	3	22	8
Vanilla LSTM	2	13	12
2-layer LSTM	4	13	5

Table 2: **Percentage of Incorrectly Predicted Labels**. BOW LR_1 denotes bag-of-words logistic regression with split across all drug classes, BOW LR_2 denotes a bag-of-words logistic regression model trained on all but one drug classes, and tested on the drug class that was left out.

One main problem with the bag-of-words logistic regression model was that it learned to associate a drug class with a label instead of picking up on the interaction between drug and food. Table 2 in the Appendix shows words with highest weights for classification with words related to drug or food compound names in red.

In addition, our logistic regression model had difficulties to classify complicated interactions in drug-food interaction sentences. In many cases, despite the same word predicting an interaction between drug and food compounds, the interaction depicted by that same word can vary depending on the context of the sentence. For example, in the sentence "Lemon verbena inhibits ACE activity", the word "inhibits" means a positive interaction between the food lemon verbena and drug ACE inhibitor. However in the sentence "Durian inhibits acetaminophen effect", the word "inhibits" means a negative interaction between the food durian and the drug `acetaminophen`.

The LSTM models can perfectly memorize the training data set reaching 98% accuracies during training. However, during testing both accuracies dropped substantially. In general, both models favored low learning rates, even in the presence of dropout regularization. However, contrary to the results of Srivastava et al. (2014), we found that a higher learning rate reduced the accuracy of the algorithms. The results were also sensitive to the choice of the optimizer, our final models were all trained with RmsProp.

Surprisingly, in spite of the limited number training examples, the LSTM implementations had higher testing accuracies where the embeddings were trained during training than the one with pre-trained embeddings (Table 1). Perhaps this is was caused by the three-year difference in the corpus; the BioNLP embeddings were trained in in 2014 while the current embeddings are based on data from 2017, and about a quarter of the food components are found only in the more recent corpus.

References

Aslam; Bushra, Rabia; Nousheen and Arshad Yar Khan. 2011. Food-drug interactions. *Oman Medical Journal*, 26(2):77–83.

Guocai Chen, Michael J Cairelli, Halil Kilicoglu, Dongwook Shin, and Thomas C Rindflesch. 2014. Augmenting microarray data with literature-based knowledge to enhance gene regulatory network inference. *PLoS Computational Biology*, 10(6):e1003666.

Adam Culbertson, Marcelo Fiszman, Dongwook Shin, and Thomas C Rindflesch. 2014. Semantic processing to identify adverse drug event information from black box warnings. In *AMIA Annual Symposium Proceedings*, volume 2014, page 442. American Medical Informatics Association.

FDA. 2005. Guidance for industry: good pharmacovigilance practices and pharmacoepidemiologic assessment. Rockville, MD: Food and Drug Administration.

Nophar Geifman, Sanchita Bhattacharya, and Atul J Butte. 2015. Immune modulators in disease: integrating knowledge from the biomedical literature and gene expression. *Journal of the American Medical Informatics Association*, 23(3):617–626.

Jessie Gerteis, David Izrael, Deborah Deitz, Lisa LeRoy, Richard Ricciardi, Therese Miller, and Jayasree Basu. 2014. Multiple chronic conditions chartbook. *AHRQ Publications*, Q14(0038).

Alan S Go, Dariush Mozaffarian, Véronique L Roger, Emelia J Benjamin, Jarett D Berry, Michael J Blaha, Shifan Dai, Earl S Ford, Caroline S Fox, Sheila Franco, et al. 2014. Heart disease and stroke statistics—2014 update: a report from the american heart association. *Circulation*, 129(3):e28.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT press.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Anne M Holbrook, Jennifer A Pereira, Renee Labiris, Heather McDonald, James D Douketis, Mark Crowther, and Philip S Wells. 2005. Systematic overview of warfarin and its drug and food interactions. *Archives of internal medicine*, 165(10):1095–1106.

Lawrence Hunter and K Bretonnel Cohen. 2006. Biomedical language processing: what's beyond PubMed? *Molecular Cell*, 21(5):589–594.

Ning Kang, Bharat Singh, Chinh Bui, Zubair Afzal, Erik M van Mulligen, and Jan A Kors. 2014. Knowledge-based extraction of adverse drug events from biomedical text. *BMC bioinformatics*, 15(1):64.

Thomas Karopka, Thomas Scheel, Sven Bansemer, and Änne Glass. 2004. Automatic construction of gene relation networks using text mining and gene expression data. *Medical informatics and the Internet in medicine*, 29(2):169–183.

Anutosh Maitra, KM Annervaz, Tom Geo Jain, Madhura Shivaram, and Shubhashis Sengupta. 2014. A novel text analysis platform for pharmacovigilance of clinical drugs. *Procedia Computer Science*, 36:322–327.

Dima M Qato, G Caleb Alexander, Rena M Conti, Michael Johnson, Phil Schumm, and Stacy Tessler Lindau. 2008. Use of prescription and over-the-counter medications and dietary supplements among older adults in the united states. *Journal of the American Medical Association*, 300(24):2867–2878.

Thomas C Rindflesch and Marcelo Fiszman. 2003. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, 36(6):462–477.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, Christopher Potts, et al. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Kenji Suzuki, editor. 2011. *Artificial Neural Networks – Methodological Advances and Biomedical Applications*. InTech, Rijeka.

Sida Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 90–94. Association for Computational Linguistics.

Liqin Wang, Peter J Haug, and Guilherme Del Fiol. 2017a. Using classification models for the generation of disease-specific medications from biomedical literature and clinical data repository. *Journal of Biomedical Informatics*, 69:259–266.

Pengwei Wang, Tianyong Hao, Jun Yan, and Lianwen Jin. 2017b. Large-scale extraction of drug–disease pairs from the medical literature. *Journal of the Association for Information Science and Technology*.

Rong Xu and QuanQiu Wang. 2013. Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. *BMC bioinformatics*, 14(1):181.

Rui Zhang, Terrance J Adam, Gyorgy Simon, Michael J Cairelli, Thomas Rindflesch, Serguei Pakhomov, and Genevieve B Melton. 2015. Mining biomedical literature to explore interactions between cancer drugs and dietary supplements. *AMIA Summits on Translational Science Proceedings*, 2015:69.

A Appendix

Relevant code in our shared Github folder, including:

- `Baseline_Random_Predict.ipynb`, contains our initial baseline
- `RNN_StanfordSentimentAnalyzer.ipynb`, contains the RNN model
- `LR_Models` and `LSTM_Models` contain our LR/LSTM models

B Appendix

		Number of sentences			
Drug	Focus	Total	Negative	Neutral	Positive
ACE inhibitors	hypertension	470	11	283	176
Acetaminophen	pain, fever	670	146	479	45
Antacids	stomach acidity	46	6	28	12
Analgesics	pain	17	1	7	9
Antihistamine	allergies	61	10	28	23
Bronchodilators	asthma	39	6	18	15
Digoxin	heart conditions	190	30	146	14
GLP-1	diabetes	222	21	130	71
Isoniazid	tuberculosis	93	19	58	16
MAO inhibitors	depression	90	8	40	42
Statins	heart conditions	357	44	242	71
Thyroxine	metabolism	212	55	121	36
	Total	2467	357	1580	530
			14%	65 %	21%

Table 3: **Label distribution of drug classes**. 2467 sentences containing negative, neutral or positive food and drug interactions were extracted from over 300,000 PubMed abstracts.

Negative	Neutral	Positive
induced	study	showed
exposure	total	enhanced
dopa	on	red
elevated	investigated	increased
protect	we	inhibitory
acetylneuraminic	studied	analgesic
while	no	monoamine
stimulated	were	which
acetaminophen	sodium	significantly
treatment	gamma	lovastatin

Table 4: **Words with highest weights for classification in Logistic Regression model**. Drug or food compound words in red

Table 5: **Classification Reports**.

Label	Precision	Recall	f1-score
Negative	0.15	0.96	0.26
Neutral	0.72	0.04	0.08
Positive	0.3	0.03	0.06
Avg	0.55	0.17	0.10

(a) **RNN sentiment analysis**.

Label	Precision	Recall	f1-score
Negative	0.69	0.39	0.50
Neutral	0.76	0.91	0.82
Positive	0.60	0.44	0.71
Avg	0.71	0.73	0.71

(b) **BOW LR$_1$**.

Label	Precision	Recall	f1-score
Negative	0.33	0.08	0.13
Neutral	0.74	0.91	0.82
Positive	0.24	0.17	0.20
Avg	0.60	0.67	0.62

(c) **BOW LR$_2$**.

Label	Precision	Recall	f1-score
Negative	0.74	0.33	0.46
Neutral	0.81	0.87	0.84
Positive	0.53	0.64	0.58
Avg	0.74	0.73	0.72

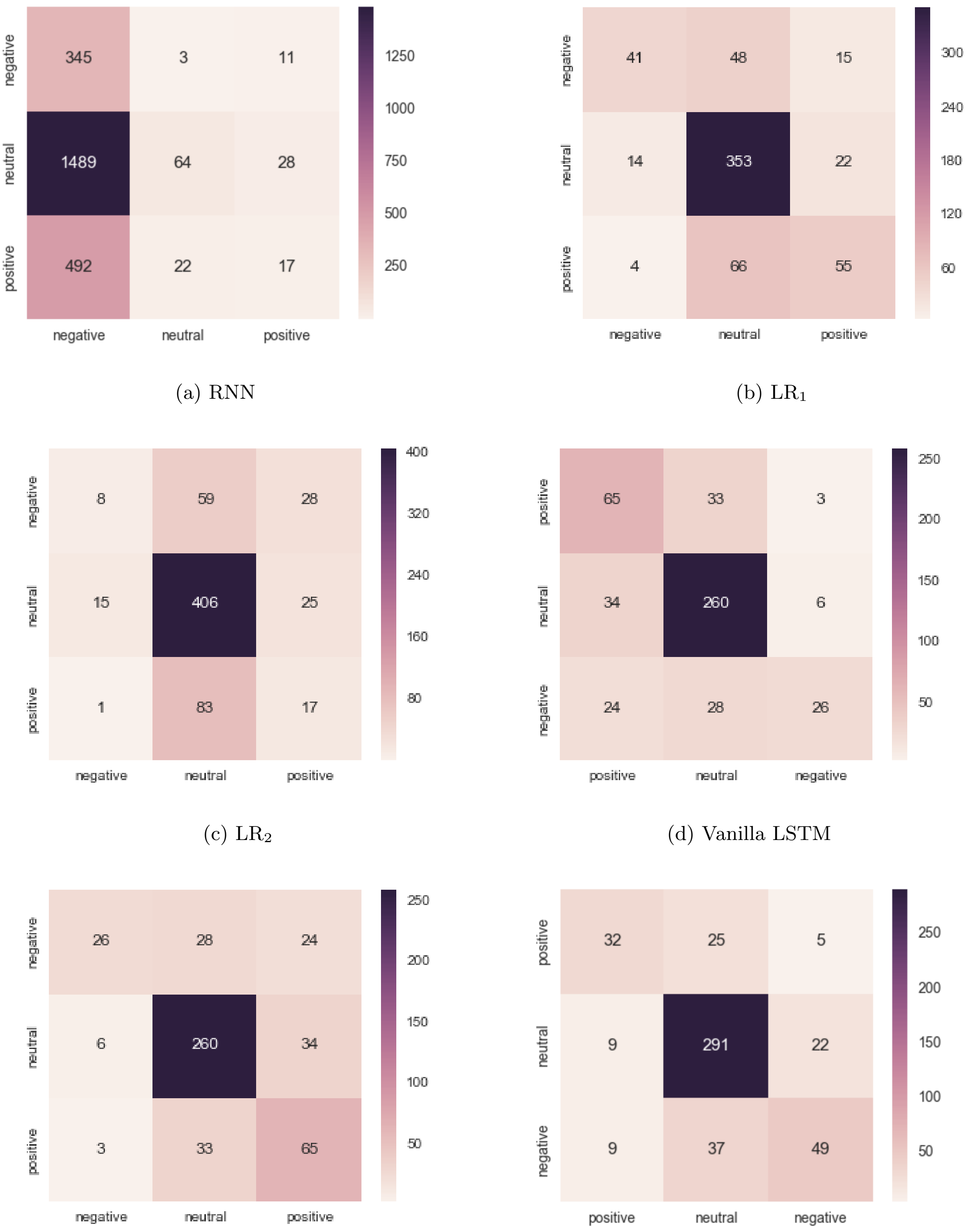
(d) **Vanilla LSTM**.

Label	Precision	Recall	f1-score
Negative	0.64	0.52	0.57
Neutral	0.82	0.90	0.86
Positive	0.64	0.52	0.57
Avg	0.76	0.78	0.77

(e) **2-layer LSTM**.

Label	Precision	Recall	f1-score
Negative	0.64	0.52	0.57
Neutral	0.82	0.90	0.86
Positive	0.64	0.52	0.57
Avg	0.76	0.78	0.77

(f) **2-layer LSTM with embeddings**.



(a) RNN

(b) LR_1

(c) LR_2

(d) Vanilla LSTM

(e) 2-layer LSTM

(f) 2-layer LSTM with pre-trained embeddings

Figure 2: Confusion matrices