SMM4H Shared Task 2020 - A Hybrid Pipeline for Identifying Prescription Drug Abuse from Twitter: Machine Learning, Deep Learning, and Post-Processing

Isabel Metzger[1,5], Emir Y. Haskovic[2], Allison Black[3], Whitley M. Yi[4], Rajat S. Chandra[1],
Mark T. Rutledge[1], William McMahon[1], and Yindalon Aphinyanaphongs[5]

[1]Sumitovant Biopharma, Inc., New York, NY
[2]Lokavant, New York, NY
[3]Roivant Sciences, Inc., New York, NY
[4]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of Colorado, CO
[5]Department of Population Health, NYU Langone Health, New York, NY
{isabel.metzger, rajat.chandra, mark.rutledge, bill.mcmahon}@sumitovant.com,
emir.haskovic@lokavant.com,
allison.black@roivant.com,
whitley.yi@ucdenver.edu,
yin.a@nyulangone.org

Abstract

This paper presents our approach to multi-class text categorization of tweets mentioning prescription medications as being indicative of potential abuse/misuse (A), consumption/non-abuse (C), mention-only (M), or an unrelated reference (U) using natural language processing techniques. Data augmentation increased our training and validation corpora from 13,172 tweets to 28,094 tweets. We also created word-embeddings on domain-specific social media and medical corpora. Our hybrid pipeline of an attention-based CNN with post-processing was the best performing system in task 4 of SMM4H 2020, with an F1 score of 0.51 for class A.

1 Introduction

Substance abuse is a major public health crisis in the United States (Substance Abuse and Mental Health Services Administration, 2019). Use of natural language processing (NLP) for automatic detection of prescription abuse or misuse in social media posts holds promise as a toxicovigilance strategy for near real-time monitoring of prescription abuse patterns and emerging trends (Hu et al., 2019; Sarker et al., 2020). Research has shown that abuse-indicating social media posts correlate with opioid-related overdose deaths at a geographical level (Sarker et al., 2019). Despite its potential, social media poses significant challenges for detection of self-reported abuse or misuse of prescriptions, including colloquial use of drug names, sparse information, and figurative language. The Social Media Mining for Health Applications (SMM4H) Task 4 challenges participants to classify tweets' mentions of opioids, benzodiazepines, atypical antipsychotics, central nervous system (CNS) stimulants or GABA analogues as potential abuse/misuse (A), consumption (C), mention-only (M) or unrelated (U) (Klein et al., 2020). As patterns of abuse and misuse can vary based on drug class and mechanism of action, our team collaborated with clinical subject matter experts (SMEs) to complete the task.

2 Data and Methods

2.1 Text Pre-processing

Publicly available tweets were pre-processed using the ekphrasis (Baziotis et al., 2017) python package to normalize url, email, percent, money, phone, user, time, and date terms and to annotate hashtags, allcaps, elongated, repeated, emphasis, and censored terms.

Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 57–62
Barcelona, Spain (Online), December 12, 2020.

2.2 Word Representations

We used fastText (Bojanowski et al., 2017) to train a 300 dimensional skip-gram model on a domain-specific corpus of over 6.5 million unique sentences for 500 epochs at a learning rate of 0.025 and a negative sampling loss. Subword character n-gram length were set between 3 and 6 (Bojanowski et al., 2017). Other parameters were set to the suggested values.[1]

Source	Number of Tweets/Sentences
Twitter Stream Archive (archive.org)	3,106,783
Tweets pulled using original drug terms listed in (O'Connor et al., 2020)	85,188
UCI Drug Review datasets (Gräßer et al., 2018)	1,313,299
Consumer Health Question Answer (CHQA) Corpus (Kilicoglu et al., 2018)	2,595
First 1 billion bytes of English Wikipedia [2]	5,244,360

Table 1: Corpora - Text Sources for Unsupervised fastText Word-Embeddings Models

2.3 Drug Lexicon Construction

A gazetteer of drug terms was produced in collaboration with SMEs to aid in feature engineering and data augmentation. We began by selecting the drug terms listed in (O'Connor et al., 2020), and supplemented these with clinically relevant drugs within the same Established Pharmacologic Classes (U.S. Food and Drug Administration, 2018), as well as with other drug classes commonly known to be misused. For each drug term, we compiled a list of its generic and proprietary names along with common misspellings. Additional synonyms and street names for the individual drugs and drug classes (e.g. "benzo" for benzodiazepine) were also identified. Generic and proprietary terms were sourced via SMEs and misspellings were sourced from (Drugs.com, 2020). Select street names were sourced from (Drug Enforcement Administration, 2020). The final gazetteer was composed of 258 terms, representing 54 drugs.

2.4 Data Augmentation

The original training and validation sets were mixed and split into new train and validation sets in a 4:1 ratio. Data augmentation was performed on the new training set with Snorkel[3], which uses a matrix completion-style modeling approach as described in (Ratner et al., 2018). Functions were created following the methods described in (Wei and Zou, 2019), including randomly replacing person names, switching adjectives, and replacing nouns, verbs, and adjectives with synonyms. Additional functions include adding or deleting a random character from tweets and replacing references to family members (e.g. swapping "brother" with "father"). Lastly, functions specific to this use case were also implemented, including replacing a drug name with one from a similar class, replacing verbs that are commonly associated with abuse (e.g switching "inject" with "shoot up"), and replacing verbs that are related to non-abusive consumption (e.g. swapping "prescribe" with "refill"). A random policy was used such that up to four of these transformation functions were applied uniformly at random per tweet if the rule was applicable.

2.5 Support Vector Machine with Feature Engineering

SVMs have been successfully applied to automated classification of text, and are commonly used as baseline for evaluating the effectiveness of new model development (Kowsari et al., 2019). For our baseline we trained an SVM using a linear kernel along with a "one-vs-rest" decision boundary scheme and the following features:

- Vader Sentiment Score[4]
- Binary flag features including presence of commonly abused drug terms, (see 2.3) and presence of emojis commonly associated with drug abuse
- Count-based features such as number of hashtags, username mentions, emojis/emoticons and number of drug terms identified (see 2.3)

[1] https://fasttext.cc/docs/en/options.html

[2] https://github.com/facebookresearch/fastText/blob/master/get-wikimedia.sh

[3] https://www.snorkel.org

[4] Vader compound score was used to measure sentiment https://github.com/cjhutto/vaderSentiment

- Co-occurrence of chemical and disease entities, where entity extraction was performed using en_ner_bc5cdr_md[5]
- tf-idf of tokens (unigrams and bigrams)

2.6 Convolutional Neural Network Architectures

Convolutional neural networks (CNNs) have shown success in short text classification, especially when orthology is important, and is computationally efficient (Kim, 2014; Minaee et al., 2020). Pre-training was performed with the following parameters (conv depth of 4, cosine loss function, drop out of 0.4, and batch size of 64) on the raw Task 4 dataset with the word embeddings from section 2.2 to build "token to vector" (tok2vec) weights. The tok2vec layer initializes the embedding matrix. The embedding layer also embeds linguistic features such as shape, suffix and prefix. We initially investigated using a 4-layer CNN with mean-pooling and softmax. In our final and winning system, we replaced the 4-layer CNN with a 2-layer hierarchical CNN and used a parametric attention layer as described in (Yang et al., 2016). The resulting weighted summary vector was then passed through a multi-layer perceptron, where each neuron was stacked with a unigram logistic regression to produce four label predicted probabilities similar to the output layer in the attention-based CNN in (Yin et al., 2015). The final classification model architecture is illustrated in figure 1. Replacing a non-linearity with a logistic regression has been shown to be effective (Zhining et al., 2016; Yin et al., 2015). We used a compounding batch size that was trained for 32 epochs. Our post-processing pipeline (2.7) was then applied to the predictions to produce the final submission.

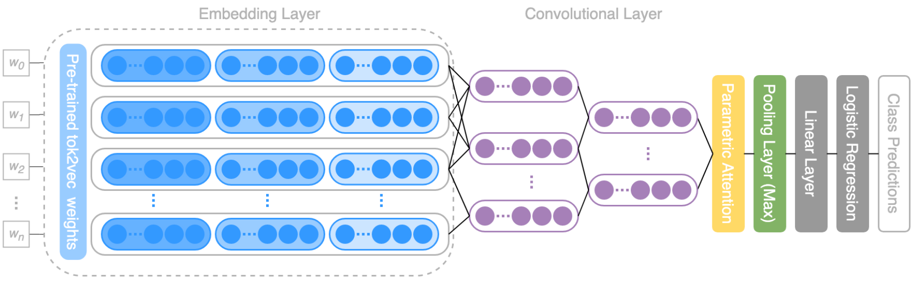


Figure 1: Attention Based CNN with Output Layer

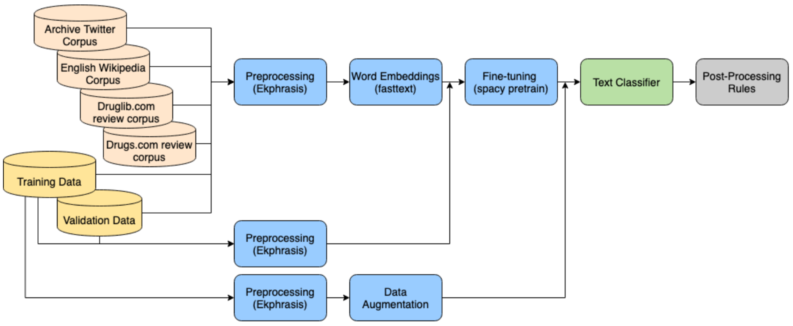
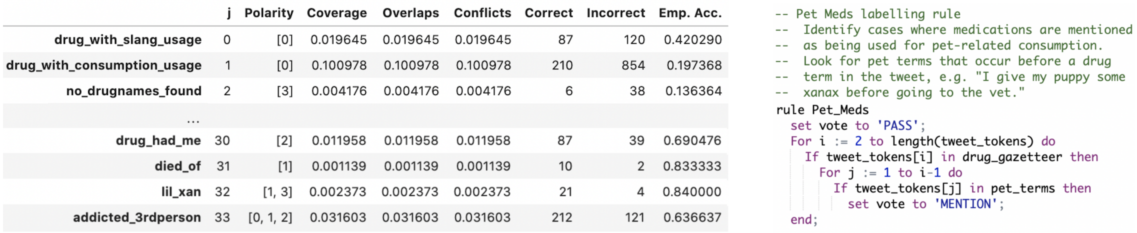

Archive Twitter Corpus
English Wikipedia Corpus
Druglib.com review corpus
Drugs.com review corpus
Training Data
Validation Data
Preprocessing (Ekphrasis)
Word Embeddings (fasttext)
Fine-tuning (spacy pretrain)
Text Classifier
Post-Processing Rules
Preprocessing (Ekphrasis)
Preprocessing (Ekphrasis)
Data Augmentation

Figure 2: Processing pipeline

2.7 Label Functions and Post-processing

Heuristic post-processing rules were created in conjunction with SMEs and transcribed into Snorkel[3] labelling functions. Rule performance was evaluated on the validation data, and a subset of rules with an empirical accuracy of 0.65 or above were selected and ordered by relative precision. These were then applied to tweets after being fed through the text classifier. For each tweet the rules are applied in succession until a rule "votes" to set the prediction label, at which point no further rules are applied and the tweet's label is updated. If no rules vote then the label provided by the classifier is unchanged.

[5] https://allenai.github.io/scispacy/



	j	Polarity	Coverage	Overlaps	Conflicts	Correct	Incorrect	Emp. Acc.
drug_with_slang_usage	0	[0]	0.019645	0.019645	0.019645	87	120	0.420290
drug_with_consumption_usage	1	[0]	0.100978	0.100978	0.100978	210	854	0.197368
no_drugnames_found	2	[3]	0.004176	0.004176	0.004176	6	38	0.136364
...								
drug_had_me	30	[2]	0.011958	0.011958	0.011958	87	39	0.690476
died_of	31	[1]	0.001139	0.001139	0.001139	10	2	0.833333
lil_xan	32	[1, 3]	0.002373	0.002373	0.002373	21	4	0.840000
addicted_3rdperson	33	[0, 1, 2]	0.031603	0.031603	0.031603	212	121	0.636637

(a) Subset of rules and their empirical scores

```
-- Pet Meds labelling rule
--  Identify cases where medications are mentioned
--  as being used for pet-related consumption.
--  Look for pet terms that occur before a drug
--  term in the tweet, e.g. "I give my puppy some
--  xanax before going to the vet."
rule Pet_Meds
  set vote to 'PASS';
  For i := 2 to length(tweet_tokens) do
    If tweet_tokens[i] in drug_gazetteer then
      For j := 1 to i-1 do
        If tweet_tokens[j] in pet_terms then
          set vote to 'MENTION';
  end;
```

(b) Example rule pseudo-code

Figure 3: Labelling functions evaluated for post-processing

3 Discussion

3.1 Error Analysis

Our hybrid pipeline commonly labeled abuse (A) tweets as mentions (M) and to a lesser degree as consumption (C). Manual review showed the model struggled to detect abuse when the subject of the abuse was of a relative or close friend, which was included as a criteria for the definition of abuse. This may be due to the low prevalence of family member- or relative-related abuse in the training dataset.

Among abuse tweets mislabeled as mention from a subset of the Task 4 validation set, the most common medication class was benzodiazepines (33/78), followed by opioids (20/78) and then CNS stimulants (19/78). Of the benzodiazepines, the most common drug among the mislabeled abuse tweets was Xanax, which could be due to its popularity in everyday vernacular, making it more challenging for an NLP model to delineate subtle differences in tweets' semantic meaning. For example, "xanax" has been used as an adjective to describe individuals or a group or used as a figure of speech to express the need to calm down (see examples below).

- `If one more entitled Xanax mom yells at me to honor a coupon that expired 3 years ago I'm jumping off a cliff`
- `honestly i feel like walkin around with xanax and just throwing it at people cause ya just needa chilll`

	precision	recall	f1-score	support
ABUSE	0.732	0.634	0.679	448
CONSUMPTION	0.781	0.858	0.817	730
MENTION	0.882	0.876	0.879	1353
UNRELATED	0.931	0.913	0.922	104

Table 2: Detailed Performance of final CNN architecture on validation split + augmented validation data

4 Results and Conclusion

In this paper, we evaluate two systems as shown in Table 3.

model system	**F1 Score**	**Precision**	**Recall**
SVM + Feature engineering	0.49	0.5276	0.4553
CNN with Attention and stacked Linear + post-processing	0.51	0.5306	0.4831

Table 3: System performance on official test data

Opportunities to improve upon current work include development of additional post-processing rules to improve labeling for edge cases and further refinement of the heuristics with the guidance of SMEs. Directions for future work in the field of NLP and prescription drug abuse include exploration of alternative methods to disambiguate ground truth labels. Due to the sparsity of contextual information available to confirm the suspicion of abuse or misuse in an individual tweet, a future strategy is to consider the creation of confidence levels for ground truth abuse/misuse labels to differentiate between high confidence of abuse versus subjective hints of abuse.

References

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. DataStories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada, August. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Drug Enforcement Administration. 2020. *Drugs of Abuse, A DEA Resource Guide*. Drug Enforcement Administration.

Drugs.com. 2020. Phonetic and wildcard search at drugs.com. `https://www.drugs.com/search-wildcard-phonetic.html`.

Felix Gräßer, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder. 2018. Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. In *Proceedings of the 2018 International Conference on Digital Health*, DH '18, page 121–125, New York, NY, USA. Association for Computing Machinery.

Han Hu, Nhathai Phan, Soon A Chun, James Geller, Huy Vo, Xinyue Ye, Ruoming Jin, Kele Ding, Deric Kenne, and Dejing Dou. 2019. An insight analysis and detection of drug-abuse risk behavior on twitter with self-taught deep learning. *Computational Social Networks*, 6(1):10, November.

Halil Kilicoglu, Asma Ben Abacha, Yassine Mrabet, Sonya E. Shooshan, Laritza Rodriguez, Kate Masterton, and Dina Demner-Fushman. 2018. Semantic annotation of consumer health questions. *BMC Bioinformatics*, 19(1).

Yoon Kim. 2014. Convolutional neural networks for sentence classification.

Ari Z. Klein, Ilseyar Alimova, Ivan Flores, Arjun Magge, Zulfat Miftahutdinov, Anne-Lyse Minard, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2020. Overview of the fifth Social Media Mining for Health Applications (#SMM4H) Shared Tasks at COLING 2020. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop Shared Task*.

Kowsari, Jafari Meimandi, Heidarysafa, Mendu, Barnes, and Brown. 2019. Text classification algorithms: A survey. *Information*, 10(4):150, Apr.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2020. Deep learning based text classification: A comprehensive review.

Karen O'Connor, Abeed Sarker, Jeanmarie Perrone, and Graciela Gonzalez Hernandez. 2020. Promoting reproducible research for characterizing nonmedical use of medications through data annotation: Description of a twitter corpus and guidelines. *J Med Internet Res*, 22(2):e15861, Feb.

Alexander Ratner, Braden Hancock, Jared Dunnmon, Frederic Sala, Shreyash Pandey, and Christopher Ré. 2018. Training complex models with multi-task weak supervision.

Abeed Sarker, Graciela Gonzalez-Hernandez, Yucheng Ruan, and Jeanmarie Perrone. 2019. Machine learning and natural language processing for geolocation-centric monitoring and characterization of opioid-related social media chatter. *JAMA Network Open*, 2(11).

Abeed Sarker, Annika DeRoos, and Jeanmarie Perrone. 2020. Mining social media for prescription medication abuse monitoring: a review and proposal for a data-centric framework. *J. Am. Med. Inform. Assoc.*, 27(2):315–329, February.

Substance Abuse and Mental Health Services Administration. 2019. Key substance use and mental health indicators in the united states: Results from the 2018 national survey on drug use and health (HHS publication no. pep19-5068, nsduh series h-54). Technical report, Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration, Rockville, MD.

U.S. Food and Drug Administration. 2018. Pharmacologic class. `https://www.fda.gov/industry/structured-product-labeling-resources/pharmacologic-class`.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. *CoRR*, abs/1901.11196.

Zichao Yang, Diyi Yang, Chris Dyer, X. He, Alex Smola, and E. Hovy. 2016. Hierarchical attention networks for document classification. In *HLT-NAACL*.

Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2015. ABCNN: attention-based convolutional neural network for modeling sentence pairs. *CoRR*, abs/1512.05193.

Lang Zhining, Gu Xiaozhuo, Zhou Quan, and Xu Taizhong. 2016. Combining statistics-based and cnn-based information for sentence classification. *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1012–1018.