Leveraging Web Based Evidence Gathering for Drug Information Identification from Tweets

Rupsa Saha, **Abir Naskar**, **Tirthankar Dasgupta** and **Lipika Dey**

TCS Innovation Lab, India
(rupsa.s, abir.naskar, dasgupta.tirthankar, lipika.dey)@tcs.com

Abstract

In this paper, we have explored web-based evidence gathering and different linguistic features to automatically extract drug names from tweets and further classify such tweets into Adverse Drug Events or not. We have evaluated our proposed models with the dataset as released by the SMM4H workshop shared Task-1 and Task-3 respectively. Our evaluation results shows that the proposed model achieved good results, with Precision, Recall and F-scores of 78.5%, 88% and 82.9% respectively for Task1 and 33.2%, 54.7% and 41.3% for Task3.

1 Introduction

Use of data generated through social media for health studies is gradually increasing. It has been found that Adverse Drug Events (ADEs) are one of the leading causes of post-therapeutic death. Thus, their identification constitutes an important challenge. Social media platforms provide significant insights about drugs usage and their possible effects, as discussed by the general public outside the controlled environment of a trial program.

The shared task offers four different subtasks, out of which we focus on two : a) Sub Task 1 : Automatic detection of posts mentioning a drug name (binary classification) and b) Sub Task 3 : Automatic classification of adverse drug reaction mentioning posts (binary classification) (Weissenbacher et al., 2018). In the following section, we briefly describe the data used to build our systems. Section 3 describes the two systems in detail, followed by the results, and a final section consisting of our observations.

2 Data Description

2.1 Task 1

The provided training set of tweet ids and labels for Task 1 listed 9623 tweets, out of which 4975 were marked with label 1 ("yes"), i.e. tweets containing mention of drug product names and/or dietary supplements. However, due to network constraints or unavailability of tweets, we could only obtain 2496 tweets. Of these, 1440 were of label "1", and 1056 were labeled "0". For the purpose of building our system, we split this set in a 85:15 ratio, and 375 tweets (216-"1", 159-"0") were used for validation, while the rest were used for building the system.

Table 1: Number of tweets available, accessed and distribution of accessed tweets across Training and Validation

	Task-1			Task-3		
	Total	Label	#tweets	Total	Label	#tweets
Provided	9623	1 0	4975 4648	25623	1 0	2224 23399
Available	2496	1 0	1440 1056	13520	1 0	1109 12411
Train	2121	1 0	1219 902	10817	1 0	888 9929
Validation	375	1 0	221 154	2703	1 0	221 2482
Test	5382			5000		

For Task 3, the provided training set of tweet ids and labels listed 25623 tweets out of which we were able access 13520. Out of these, a mere 1109 were labeled as tweets containing mention of adverse drug events. We divided this set in a 80:20 ratio, with 2703 tweets (221- "1", 2482 - "0") as the validation set and the remaining 10817 for training.

3 System Description

3.1 Task 1 : Automatic detection of posts mentioning a drug name

As preprocessing, each tweet is tokenized and tagged using the Ark-Tweet-NLP tool (Owoputi et al., 2012). From the resultant tokens, we considered those that are tagged as Proper Nouns or Common Nouns. Such tokens are passed to the Information Gathering Module.

Proceedings of the 3rd Social Media Mining for Health Applications (SMM4H) Workshop & Shared Task, pages 67–69
Brussels, Belgium, October 31, 2018.

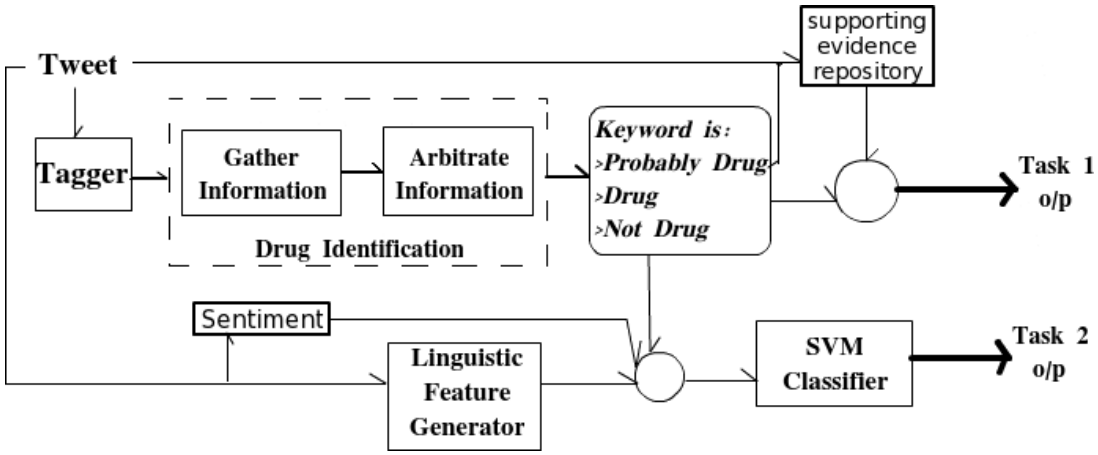

Figure 1: Overview of architecture

Table 2: Preliminary results from internal evaluation on the training dataset

Approach	Class	Precision	Recall	F-Score	Accuracy
Task 1	Class 0	0.58	0.70	0.63	
	Class 1	0.70	0.57	0.63	0.63
	Wt. Avg	0.64	0.63	0.63	
Task 3 (1)	Class 0	0.96	0.8	0.88	
	Class 1	0.23	0.64	0.34	0.79
	Wt. Avg	0.89	0.79	0.83	
Task 3 (2)	Class 0	0.95	0.85	0.9	
	Class 1	0.27	0.58	0.37	0.83
	Wt. Avg	0.9	0.83	0.86	

This module retrieves information relevant to the keyword from three different sources : Wikidata (Vrandečić and Krötzsch, 2014), Wikipedia data dumps (Wu and Weld, 2010) and Wordnet (University, 2010). The module searches for evidence that a word represents a drug/supplement, in the corresponding gloss, hierarchy structure, and web page structure, as obtained from each source. Wikidata is the source of structured information, and presence or absence certain keys (e.g. RxNorm Id.,drug interaction etc.) are used as evidence. On the other hand, Wikipedia is mostly unstructured textual information. From this source, evidence may be found in the form of the definition of the keyword, the presence of "side effects" of the keyword, the hierarchical category the entry belongs to, among other ways. From Wordnet, we use both the gloss and the hierarchy structure.

The obtained information is further fed to the Information Arbitration Module. The Arbitration module considers the different information obtained with regard to a particular keyword, and returns a judgment as to whether the the keyword is "Not Drug", "Probably Drug " or "Drug".

In case a keyword receives a "Probably Drug" judgment, it may be a drug name depending on the information obtained from neighbouring tokens. In such cases, we extract the most frequent keywords co-occurring with the keywords to create a repository of terms. this supporting evidence repository contains a collection of patterns, obtained from the training set, which dictate under which conditions a "Probably Drug" keyword can be upgraded to a "Drug". e.g. "Protein" by itself is not a supplement name, however, "Protein shakes" is, when used as treatment.

3.2 Task 3 : Automatic classification of adverse drug reaction mentioning posts

For classification, we employ a SVM based classifier with a polynomial kernel (Dasgupta et al., 2017). The features used are: *(a) PMI*: the Pointwise Mutual Information (PMI)(Bouma, 2009) between all possible bigram pairs are considered. Co-occurrence is counted at the sentence level, i.e. *P(i, j)* is estimated by the number of sentences that contain both terms W_i and W_j , and *P(i)* and *P(j)* are estimated by counting the total sentences containing W_i and W_j , respectively. Only those bi-grams whose PMI score exceeds the *average* + *stddev* threshold, are retained as features, *(b) Term Relevance*: all unigram terms that are relevant to

Table 3: Results of Experiments on Final Test Set

Task		Precision	Recall	F-Score
Task 1	Team ART	0.79	**0.88**	0.83
	Task Average	**0.89**	0.87	**0.88**
Task 3	Team ART (1)	0.305	**0.627**	0.411
	Team ART (2)	0.332	0.547	**0.413**
	Task Average	**0.39**	0.52	0.40

the positive class, *(c) Dependency feature counts*: counts of all Stanford typed dependency features, and *(d) Drug Name*: The drug names present in the tweet, as obtained employing the same Drug Identification Module mentioned in 3.1. Since the data is heavily skewed in favour of negative examples, we train a total of 11 models, each with a non-skewed subset of the data. The training data for each model consists of 909 positive and approximately 925 negative examples, with negative examples, sampled randomly. For each test data, each of the 11 models predict the "yes"/"no" label, and all the predictions are fed to an arbitrator for a final decision.

As an enhancement, we also use the sentiment polarity score as an additional feature, using the VADER sentiment analysis tool (Gilbert, 2014). Using sentiment does result in a performance improvement, as noted in section 4.

4 Results and Observations

The results as obtained from splitting the initial data into training and test sets are tabulated in Table 2. For Task 1, we report the Precision, Recall, F-Score and Accuracy on the whole of the training set. However, for Task 3, since the data is skewed, we report all three versions (for negative class, for positive class and for weighted average of both classes) of these same parameters. Task 3 (1) represents the results for experimentation without using sentiment polarity, and Task 3 (2) are experiments with the sentiment factor.

The results on the final test set are reported in Table 3. We compare our result with the mean of results obtained by other participating teams. For Task 3, only the results with respect to the positive class was from others were available.

The results for Task 1 are poor because, while the Drug Identification module is successful in pointing out keywords which are drugs/supplements in the tweet, it does not have the capability to distinguish whether that keyword is used to imply medication or not. For example, in *"I wanna name my first child vyvanse"*, while *"vyvanse"* is a drug, here it is clearly not used in a medication sense of the term, and the given label is 0. Our method fails in such cases.

Results for Task 3 may benefit on using a more robust sentiment feature scorer, especially one that is trained on drug tweets themselves. We can also use different classification methods to test if results improve further.

References

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40.

T Dasgupta, A Naskar, and L Dey. 2017. Exploring linguistic and graph based features for the automatic classification and extraction of adverse drug effects. In *18th International Conference on Computational Linguistics and Intelligent Text Processing*.

CJ Hutto Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf.*

Olutobi Owoputi, Brendan OConnor, Chris Dyer, Kevin Gimpel, and Nathan Schneider. 2012. Part-of-speech tagging for twitter: Word clusters and other advances. *School of Computer Science*.

Princeton University. 2010. Wordnet. *About WordNet*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

D Weissenbacher, A Sarker, M Paul, and G Gonzalez-Hernandez. 2018. Overview of the third social media mining for health (smm4h) shared tasks at emnlp 2018. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Fei Wu and Daniel S Weld. 2010. Open information extraction using wikipedia. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 118–127. Association for Computational Linguistics.