International Journal of Future Generation Communication and Networking
Vol. 13, No. 1s, (2020), pp. 226-- 233

Argument Mining for Medical Reviews

Abhiruchi Bhattacharya[1], Kasturi Kumbhar[2], Padmaja Borwankar[3*], Ariscia Mendes[4], Sujata Khedkar[5]
Department of Computer Engineering,
V.E.S. Institute of Technology,
Chembur, India

Abstract
Argument mining is the process of extracting opinions and reasons from dialectical text and drawing conclusions to illuminate the author's viewpoint concisely. Hence, argument mining becomes highly useful in the medical domain, especially for pharmacists and analysts in analysing the effects of drugs on people and their varying opinions on the effectiveness of the drugs in question. In this paper, we propose a system that uses argument mining and machine learning to extract supporting and attacking relationships between sentences from drug reviews, in an effort to build an application that can provide deeper insight into people's opinions on various drugs. We identify argumentative content based on the presence of discourse indicators, which then undergoes pre-processing and feature extraction to form a meaningful representation of the text. We consider seven feature sets consisting of structural features, TF-IDF scores for unigrams and bigrams and their combinations. The feature vectors are given to a machine learning classifier for predicting support/attack relations between sentence pairs. We evaluate three classification algorithms, namely support vector machine, random forest classifier and AdaBoost classifier, using precision, recall, F1 scores and 10-fold cross validation accuracy as evaluation parameters. The application can then give a detailed analysis of the given medical review.
Keywords: *Argument mining, Drug reviews, Pharmacovigilance, Natural language processing, Machine learning, Relationship extraction.*

1. Introduction

Due to the widespread prevalence of the internet, many people now have a common medium to share both technical information and casual opinions about a variety of subjects. This provides a rich field of study for natural language processing research, and there is a need to provide a summarized view of the huge amount of data and extract meaningful information from it.

Developing and applying computational models of argument is very important for the healthcare domain. Healthcare information is complex, heterogeneous and inconsistent. With the advent of new drugs in the market, there is a decent chance that someone might suffer from an adverse drug reaction unforeseen by the manufacturers of the drug. Adverse reactions are the recognized hazards of drug therapy and they can occur with any class of drugs. Tracking the discourse happening about medicines and identifying it from medical reviews becomes important in order to better understand the effects of drugs on the human body, apart from clinical trials. Argument mining is appealing for medical reviews as it allows for important conflicts to be highlighted and analyzed and unimportant details to be suppressed. The general public's reception of drugs and to take precautionary measures while prescribing such drugs can be analyzed more effectively by developing automatic argument mining techniques.

Hence, we aim to build a system that can extract supporting or attacking relationships between arguments from drug reviews using machine learning and natural language processing techniques. These arguments should then be presented so as to support sense-making of the target domain. By looking at the arguments related to a medical topic or a drug, both medical professionals and general users can understand the general reception and opinions of the public

ISSN: 2233-7857 IJFGCN

and the reasons and evidence behind these opinions, thereby better equipping themselves to make sound decisions about medical prescriptions and intakes.

2. Literature Survey

Previous work related to creating search engines for argumentative content include IBM's Project Debater [1-4]. Project Debater involved the creation of an AI-based debate opponent, for which it became necessary to construct a search engine framework to fetch arguments related to a particular topic. Levy et. al. [1] use a supervised learning model to extract context dependent claims related to specific topics from a set of documents. A series of logistic regression classifiers and maximum likelihood probability models are used to sort through the dataset and extract claims from within sentences. Rinott et. al. [2] developed an architecture based on a similar pipeline of modular components to extract context dependent evidence from a set of topics and a claim. Levy et. al. [3, 4] expand the concept of automatic claim extraction to create an unsupervised claim detector. In [3], the assumption is that the main concept of a sentence generally occurs after the word 'that'. A claim within such a sentence is detected based on the presence of one or more words from a 'claim lexicon', which contains some typical words that mark the beginning of an opinion or claim in standard English. Word2Vec embeddings (Mikolov et. al. [5]) are used to detect the presence of a main concept. This idea is extended in [4] to include more flexibility by considering sentences that do not conform strictly to the syntax in [3]. [4] uses two Bi-LSTM neural networks, one trained on sentences including the word 'that' and the other on sentences that do not include 'that'. The coverage of such systems is low, however, as a claim can be worded in many ways, especially so when considering informal language used in user reviews.

Stab et. al. [6] use a variety of feature sets (structural, lexical, syntactic, contextual and indicators) to identify argumentative structure in essays. The authors use a manually annotated corpus of clauses taken from persuasive essays, connected by support and attack labels. The tasks of detecting argumentative content and extracting relations are treated as two separate classification problems. The former is treated as a multiclass classification problem of labelling a clause as a major claim, claim, premise or neither. Relation identification is treated as a sentence pair classification problem with two classes, namely 'support' and 'non support'. Best performance is obtained using SVM and lexical and syntactic features are reported as the most influential feature sets, along with structural features for sentence pair classification. Aker et. al. [7] rigorously explore the effectiveness of all combinations of the feature sets defined in [6], with the addition of word embeddings. The authors test these combinations on persuasive essays as well as the Wikipedia corpus presented in Aharoni et al.[8]. Structural features are found to be the most robust for both identification and relationship extraction tasks, with Random Forest classifier generally performing well for both datasets.

Lawrence and Reed [9] combine three machine learning approaches for argument mining. They use discourse indicators as a means to identify a connection between arguments, then use argument schemes and topic similarity to connect those propositions that remain unconnected after the initial step.

Cocarascu et. al. [10] explores the mining of attack and support relations between two sentences. The initial GloVe [11] embeddings of the two sentences are given to two parallel deep neural nets to form the independent vector representations of both sentences. These are combined (by either summing or concatenating) and given to a softmax classifier to finally give the support, attack or unrelated label between the two sentences. These labels are further used to extract bipolar argumentation frameworks from hotel reviews and measure the dialectical strength of a review, which, along with other features, are given to a random forest classifier to detect deceptive hotel and restaurant reviews. While complicated, this provides a promising architecture for detecting spam or contradictory reviews.

ISSN: 2233-7857 IJFGCN

Chawla et al. [12] propose an analytical tool for evaluating the effectiveness of drugs and monitoring adverse drug reactions from online drug reviews. They train a classifier to label any sentence as 'effective', 'ineffective', 'adverse' or 'none', and use its predictions to gain insights into medical reviews. They explore eight feature sets including tf-idf vectors, VADER sentiment scores, unigrams and bigrams etc. They obtain best performance using tf-idf vectors and VADER sentiment scores on a OnevsRest classifier.

Most previous work includes manual annotation of argumentative components within sentences which does not make it scalable. We use semi-automated labelling scheme where we label the relationship between a pair of sentence as support or attack based on discourse indicators.

Previous research in argument mining considers dialectical text which includes opinions on variety of topics like social issues, hotel reviews, etc. We focus on a domain specific application of argument mining to medical reviews. Reviews are written informally. Some technical terminology is used but structure and language varies widely. Hence the application can be a decision support system but not a substitute for actual prescriptions.

3. Methodology

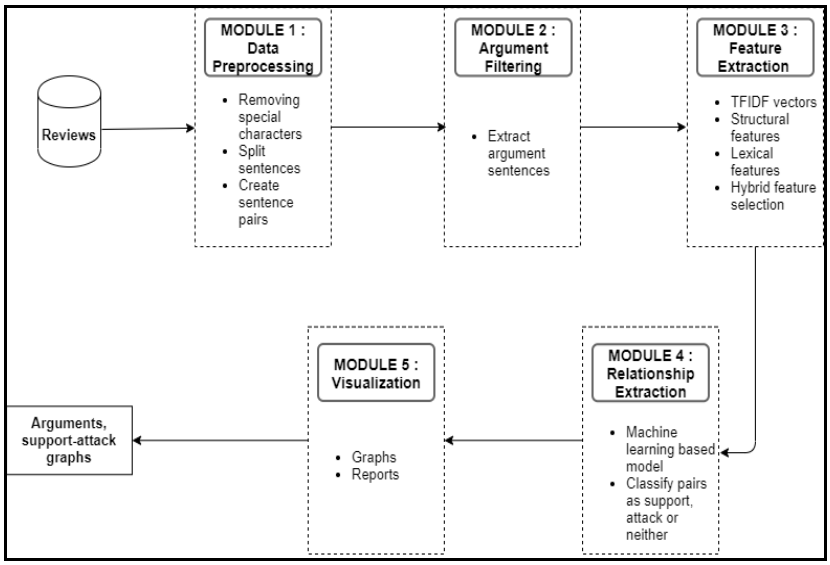

Figure 1. Modular diagram

3.1. Data collection

For our system, we use the corpus prepared in [12] which contains sentences from drug reviews for various drugs prescribed for neurological conditions. The reviews have been scraped from three websites, namely webmd.com, everydayhealth.org and drugs.com.

3.2. Preprocessing

Each review sentence is broken into phrases based on the occurrence of discourse indicators i.e. common English words that denote changes in flow of logic. Examples of discourse indicators include connectors like but, because, when etc. The next step is forming pairs of these clauses. From a total of 852 considered reviews, we form 4253 sentence pairs.

3.3. Argument Filtering

We filter argumentative sentence pairs from those obtained in the previous step by retaining those pairs with at least one discourse indicator in either sentence. For training models,

ISSN: 2233-7857 IJFGCN

labelling each of the pairs is necessary. They are categorized mainly as support, attack or neither denoted by ‘s’, ‘a’ and ‘n’ respectively. A sentence pair (s1,s2) labelled as ‘s’ or support implies that there is a supporting relationship between the constituent sentences, s1 and s2. Similarly, a pair labelled as ‘a’ or ‘n’ means that there is an attacking or neutral relationship between the pair respectively.

We adopt a semi-supervised approach to data labelling, by using both script-based and manual annotation. For script-based annotation, if the second clause contains a support indicator, then the sentence pair is labelled as 's' whereas if it contains an attack indicator, it is labelled as ‘a’. The remaining pairs are labelled manually as ‘s’, ‘a’ and ‘n’. After preprocessing and filtering from 4253 initial sentence pairs, we obtain a total of 1282 labelled sentence pairs in this manner.

3.4. Feature extraction

For feature extraction, we first consider the structural features used in [6], since structural features were found to be significant for relation extraction tasks, as reported in [7]. For a sentence pair (s1,s2), structural features comprise of

1. Number of tokens in s1
2. Number of tokens in s2
3. Absolute difference between 1 and 2
4. Number of punctuation characters in s1
5. Number of punctuation characters in s2
6. Absolute difference between 4 and 5

We also consider TF-IDF vectors as used in [12]. We first create a vectorizer using all sentences combined and extract top 15000 features (considering unigrams and bigrams). This vectorizer is used to create tf-idf vectors for both clauses in a sentence pair. Considering the six structural features and 15000 features for both sentences, our final dataset has 1282 pairs of sentences and 30006 features.

Table 1. Distribution of dataset according to category

Category	Number of pair
Neutral	59
Support	489
Attack	734

Hence, we create the following feature sets for relationship extraction:

F1: Structural features

F2: TF-IDF top 4698 only unigram features

F3: Structural features + TF IDF 4698 unigram features

F4: TF-IDF top 15000 only bigram features

F5: Structural features + TF IDF 15000 bigram features

F6: TF-IDF top 15000 unigram and bigram features

F7: Structural features + TF IDF 15000 unigram and bigram features

3.5. Relationship extraction

ISSN: 2233-7857 IJFGCN

We consider the problem of relationship extraction as a multiclass classification problem, i.e. any sentence pair can be classified into one of three classes: support ('s'), attack ('a') or neither/neutral ('n'). For training, we use 80:20 split for training and cross validation. For our initial experimentation, we consider the following classification algorithms:

1. Support Vector Machine
2. Random Forest classifier
3. Adaboost classifier

4. Results

With ten-fold cross validation, the best accuracy was found using Adaboost classifier for all three categories using structural and TF-IDF features for unigrams and bigrams, with a mean accuracy of 94.13% and standard deviation of 1.56. The SVM performed better than the Random Forest classifier, with a mean accuracy of 89.71%, and standard deviation of 1.86.

We observe that best performance is achieved by the AdaBoost classifier for all three categories using structural and TF-IDF features for unigrams and bigrams. Despite this, most misclassifications are observed for the 'n' category. It can be seen from the confusion matrices that both SVM and random forest classifiers performed poorly at identifying neutral relations between sentence pairs. The Adaboost classifier performed better, but still predicted most neutral examples to be supporting examples. This can be attributed to the dataset being skewed, with much fewer 'n' examples than 'a' or 's'.

Table 2. Model performances

Classifie	eature s	Category 'neutral		Category 'suppor		Category 'attack		Average F1	ccuracy (
		precisio	recall	precisio	recall	precisio	recall	(weighted)	fold CV)
SVM	F1	0.00	0.00	0.49	0.30	0.63	0.83	0.56	50.24%
	F2	0.00	0.00	0.48	0.43	0.64	0.72	0.56	56.67%
	F3	0.00	0.00	0.53	0.51	0.67	0.73	0.60	56.92%
	F4	0.00	0.00	0.48	0.14	0.61	0.92	0.51	46.56%
	F5	0.00	0.00	0.47	0.28	0.62	0.82	0.54	51.38%
	F6	0.25	0.09	0.92	0.85	0.89	0.97	0.88	**89.77**%
	F7	0.25	0.09	0.94	0.86	0.89	0.98	0.89	89.71%
andom Fo	F1	0.00	0.00	0.37	0.36	0.57	0.59	0.48	49.22%
	F2	0.00	0.00	0.53	0.28	0.62	0.85	0.56	52.41%
	F3	0.00	0.00	0.58	0.36	0.64	0.85	0.59	53.63%
	F4	0.00	0.00	0.43	0.06	0.60	0.96	0.47	44.01%
	F5	0.00	0.00	0.27	0.10	0.56	0.83	0.45	47.77%
	F6	0.00	0.00	0.83	0.77	0.82	0.92	0.81	**85.24**%
	F7	0.00	0.00	0.84	0.86	0.90	0.95	0.86	84.63%
AdaBoos	F1	0.00	0.00	0.68	0.26	0.63	0.92	0.58	48.81%
	F2	0.12	0.09	0.45	0.14	0.60	0.89	0.50	46.38%
	F3	0.00	0.00	0.43	0.19	0.59	0.84	0.50	47.28%
	F4	0.12	0.09	0.45	0.14	0.60	0.89	0.50	46.38%
	F5	0.00	0.00	0.62	0.05	0.59	0.98	0.47	48.36%
	F6	0.36	0.45	0.94	0.92	0.98	0.97	0.93	94.11%
	F7	0.21	0.27	0.93	0.90	0.99	0.99	0.93	**94.13**%

ISSN: 2233-7857 IJFGCN

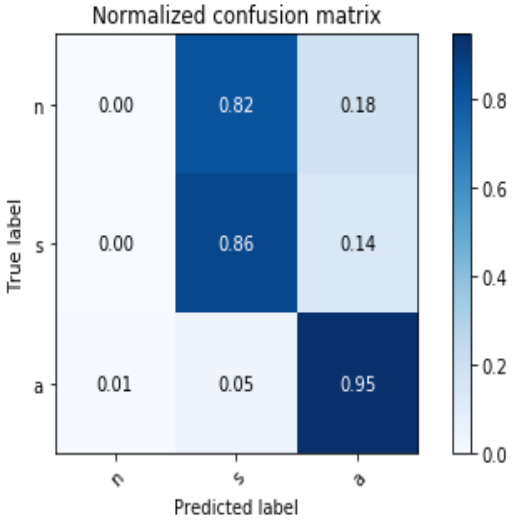

Figure 2. Random Forest confusion matrix for feature set F7

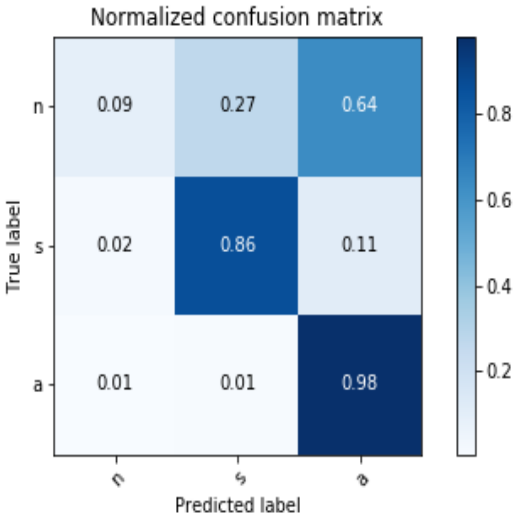

Figure 3. SVM confusion matrix for feature set F7

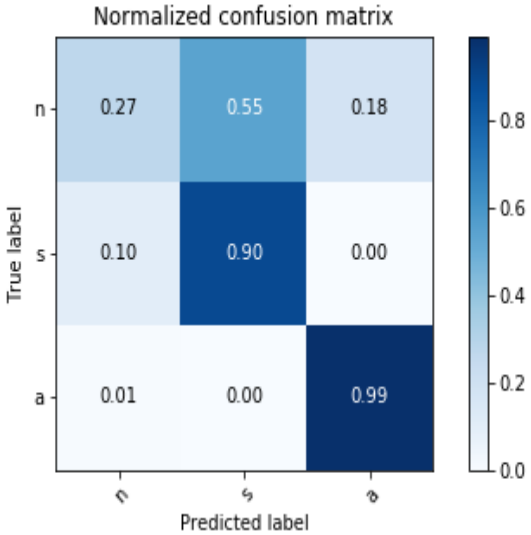

Figure 4. AdaBoost confusion matrix for feature set F7

ISSN: 2233-7857 IJFGCN

5. Conclusion

Argument Mining is the automatic identification and extraction of the structure of inference and reasoning presented in natural language. This helps to determine what opinions people have about certain topics, why they have those opinions and hence provide valuable insights in various domains. The models we trained, achieved a reasonable accuracy for support and attack relations. We achieved a maximum accuracy of 89.77% by our SVM model and 85.24% by our Random Forest model, for unigram and bigram features. The AdaBoost model performed the best with an accuracy of 94.13% for structural, unigram and bigram features. Performance can be improved by extending the dataset or data augmentation. The model needs to be tested for unseen examples to check whether it generalises. Visualization can be done by creating support/attack graphs to provide useful insights to the users. Finally, a Flask application can be designed for the frontend interface. In the future, deep learning methods such as LSTM networks can be explored which contains memory cells to store relevant information.

The reasons and evidence extracted by such a system can give doctors an indication towards potentially unknown factors behind working of the drug. It can give a deeper insight into what people actually think of a particular drug and based on it, new information about the drug can be obtained which can be used in a more constructive way in the future. This way, it can benefit both, the doctors for prescribing most appropriate drugs as well as patients for finding out novel details about the drug.

References

1. R. Levy, Y. Bilu, D. Hershcovich, E. Aharoni,& N. Slonim. “Context dependent claim detection”, In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp. 1489-1500, 2014.

2. R. Rinott, L. Dankin, C. A. Perez, M. M. Khapra, E. Aharoni, and N. Slonim, "Show me your evidence-an automatic method for context dependent evidence detection." In Proceedings of the 2015 conference on empirical methods in natural language processing, 2015, pp. 440-450.

3. R. Levy, S. Gretz, B. Sznajder, S. Hummel, R. Aharonov and N. Slonim, "Unsupervised corpus–wide claim detection", Proceedings of the 4th Workshop on Argument Mining, 2017. Available: 10.18653/v1/w17-5110.

4. R. Levy, Y. Bilu, D. Hershcovich, E. Aharoni,& N. Slonim. “Towards an argumentative content search engine using weak supervision”, In Proceedings of the 27th International Conference on Computational Linguistics, pp. 2066-2081, 2018.

5. T. Mikolov, I. Sutskever, K. Chen, G. Corrado & J. Dean. “Distributed Representations of Words and Phrases and their Compositionality.” Advances in Neural Information Processing Systems. 26. 2013.

6. C. Stab, and I. Gurevych. "Identifying argumentative discourse structures in persuasive essays." In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 46-56. 2014.

7. B. Aker, A. Sliwa, Y. Ma, R. Lui, N. Borad, S. Ziyaei, and M. Ghobadi. "What works and what does not: Classifier and feature analysis for argument mining." In Proceedings of the 4th Workshop on Argument Mining, pp. 91-96. 2017.

ISSN: 2233-7857 IJFGCN

8. E. Aharoni, A. Polnarov, T. Lavee, D. Hershcovich, R. Levy, R. Rinott, D. Gutfreund, and N. Slonim. “A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics”. In Proceedings of the First Workshop on Argumentation Mining. pages 64–68. 2014.

9. J. Lawrence, and C. Reed. "Combining argument mining techniques." In Proceedings of the 2nd Workshop on Argumentation Mining, pp. 127-136. 2015.

10. O. Cocarascu and F. Toni, "Combining Deep Learning and Argumentative Reasoning for the Analysis of Social Media Textual Content Using Small Data Sets", Computational Linguistics, vol. 44, no. 4, pp. 833-858, 2018. Available: 10.1162/coli_a_00338.

11. J. Pennington, R. Socher & C. Manning. “Glove: Global Vectors for Word Representation.” EMNLP. 14. 1532-1543. 2014. 10.3115/v1/D14-1162.

12. D. Chawla, D. Mohnani, V. Sawlani, S. Varma and S. Khedkar, "Drug review analytics of neurological disorders," 2019 International Conference on Nascent Technologies in Engineering (ICNTE), Navi Mumbai, India, 2019, pp. 1-3.