Analysis of Medical Arguments from Patient Experiences Expressed on the Social Web

Kawsar Noor[1], Anthony Hunter[1] and Astrid Mayer[2]

[1] Department of Computer Science, University College London, UK
[2] Department of Oncology, Royal Free London NHS Foundation Trust, UK

Abstract. In this paper we present an implemented method for analysing arguments from drug reviews given by patients in medical forums on the web. For this we provide a number of classification rules which allow for the extraction of specific arguments from the drug reviews. For each review we use the extracted arguments to instantiate a Dung argument graph. We undertake an evaluation of the resulting argument graphs by applying Dung's grounded semantics. We demonstrate a correlation between the arguments in the grounded extension of the graph and the rating provided by the user for that particular drug.

1 Introduction

Evidence based medicine stipulates that patients are offered medication and treatment based on scientific evidence published in the medical literature. Whilst patients may find it difficult to relate to medical statistics they are keen to understand benefits, potential side effects and implications on their life and life style. Drug reviews, much like other product reviews on the internet, provide useful insights into the performance and acceptance of the drug amongst patients who have experience of it [2]. Drug review websites contrast with traditional medical resources by providing access to an interesting set of arguments based on personal experiences of the patients. Whilst this reflects the subjective experience of individuals we propose to view the review process as users providing arguments and counter arguments about the drug in question.

If such arguments can be retrieved from drug review websites, it is possible to arrange them using existing argument-theoretic frameworks such as Dung's argument graph [4]. The generation of a Dung graph to represent the arguments in a single drug review, enables one to elicit the overall assessment of the drug based on the evaluation of the argument graph; such evaluations can be achieved using Dung's extensions. In order to validate this assessment it is possible to exploit the rating function provided by drug review websites, which enables users to numerically score the drug. We propose that by correlating the rating, produced by our argument extraction and analysis system, against the numerical rating data given by the drug review author we can ascertain a general measure as to how accurate our analysis was.

We believe this work is a novel contribution because it shows how Dung's approach to analysing arguments is reflected in the way drug review authors

evaluate conflicting arguments within a single drug review. This suggests that we could extend the application of our method to those drug review sites that do not have user provided ratings in order to generate analogous ratings. Furthermore our argument-based analysis could provide structured information to patients who are trying to garner an understanding of how the drug was received by previous users. We expect that this tool will provide patients with supplementary reasons for and against the treatment.

Note our method of extracting arguments is not meant as a contribution to argument mining, rather it is a simple method to automate the process of instantiating argument graphs and could potentially be improved by harnessing more advanced argument mining techniques such as those reviewed in [9].

2 Argument Extraction

In the following, we show how simple rule-based information extraction techniques can be harnessed to extract arguments. The implemented system has been written in Python, and makes use of the natural language processing toolkit NLTK[3]. The code and datasets are available on Github[4].

We take reviews from two medical websites (Drugs.com and Webmd.com). Drug reviews on these websites, much like other products tend to focus on a core set of features of the product. We identify a set of common features found across the various reviews. The recurrent themes tend to be centred around the side effects experienced, the overall success of the drug and the general experience with the drug.

> *"I get* $\underline{achy}_{side\ effect}$ *in the hands and feet, have* $\underline{gained\ weight}_{side\ effect}$ *(20)lbs. and* $\underline{hate}_{negative\ experience}$ *the hunger it seems to give me cravings for calorie laden foods."*

As can be seen in review above the user's focus is on the side effects of the drug, whilst some words such as 'hate' would indicate that the user had a negative experience with the drug. Similar observations were made when reading a range of different drug reviews. With these observations in mind we identified the following core themes which we use to extract arguments for/against a number of drugs: (1) Presence of side effects; (2) Severity of the side effects; (3) Polarity of experience with the drug; (4) Whether or not supplementary drugs can be taken for side effects from the primary drug.

Each theme is identified through the appearance of key words. Using the example of the theme *presence of side effects*, statements pertaining to this theme are identifiable when a side effect is mentioned; vocabulary for which can be sourced from medical literature. Furthermore each theme can be assessed for polarity, so continuing the example of the *presence of side effects* theme we say that the resulting argument types are *the absence of a side effect* and *the*

[3] http://www.nltk.org/
[4] https://github.com/robienoor/NLTKForumScraper

presence of a side effect. These argument types thus either favour or oppose the use of that particular drug. Using this approach we formalised 10 classification rules based on the themes mentioned above.

In this paper we assume that each argument is presented in a single sentence. A sentence may convey multiple arguments but no argument requires multiple sentences to convey it. This is a simplifying assumption that we do not further investigate in this paper. The role of the classification rules is to identify the types of argument present in each sentence.

In order to define the classification rules we compiled a number of lists namely `Symptoms`, `Drugs`, `Diseases`, `PosWords`, `NegWords`, `Inverters` and `SideEffects`. The list `SideEffects` contains the term *side effect* in various forms eg: *symptoms*, *side-effects* etc. The list `Inverters` contains a list of negating words eg: *no*, *not*, *none* etc. These lists serve the purpose of providing quick access to medical and sentiment terminology.

The classification rules below are formalised using first-order logic. Below is a list of predicates that are common across the classification rules.

- `Occur(sentence, wordlist, position)` which holds when there is a word in `wordlist` that occurs at the point `position` in `sentence`
- `ImmediatelyBefore(string1, string2)` which holds when `string1` is the sentence immediately before `string2`.
- `Contains(sentence, wordlist)` which holds when at least one of the words in `wordlist` is in `sentence`.
- `ArgumentType(sentence, type)` which holds when the `sentence` is of type `type`.
- `Score(sentence, wordlist)` is a function that returns the number of words in `wordlist` that occur in `sentence`

With the common predicates defined above we proceed to define all of the individual classification rules. Essentially each rule classifies a sentence to be of a particular type if the conditions of the rule are met for the sentence. A sentence may be classified to be of more than one type (though in practice this is infrequent).

1. **NoSideEffectsI**: This rule looks for an inverter word immediately followed by a side effect string.
   eg: *I have* $\underline{no_{inverter}}$ $\underline{side\ effects_{sideEffect}}$

   $\forall$`sentence, string1, string2`
   `Contains(string1,Inverters)` $\wedge$ `Contains(string2,SideEffects)`
   $\wedge$ `ImmediatelyBefore(string1,string2)`
   $\rightarrow$ `ArgumentType(sentence,noSideEffectsType1)`

2. **NoSideEffectsII**: This looks for an inverter word before a side effect string irrespective of its position in the sentence.
   eg: *During the time I took the medication I did* $\underline{not\ _{inverter}}$ *experience any* $\underline{side\ effects_{sideEffect}}$ *at all*

$\forall$sentence, position1, position2
Occur(sentence,Inverters,position1)
$\wedge$ Occur(sentence,SideEffects,position2)
$\wedge$ position1 $<$ position2
$\rightarrow$ ArgumentType(sentence,noSideEffectsType2)

3. **SideEffectsI**. This looks for a side effect string with no inverter words in the preceding words.
   eg: *The side effects$_{sideEffect}$ outweighed the good*

   $\forall$sentence, position1
   Occur(sentence,SideEffects,position1)
   $\wedge \neg\exists$position2(
   Occur(sentence,Inverters,position2)
   $\wedge$ position1 $>$ position2)
   $\rightarrow$ ArgumentType(sentence,sideEffectsPresentType1)

4. **SideEffectsII**. This searches for a symptom within a sentence.
   eg: *The side effects were gradual at first but now they are full blown...fatigue$_{symptom}$ and joint pain$_{symptom}$*

   $\forall$sentence
   Contains(sentence,Symptoms)
   $\wedge \neg$Contains(sentence,PosWords) $\wedge$ $\neg$Contains(sentence,NegWords)
   $\rightarrow$ ArgumentType(sentence,sideEffectsPresentType2)

5. **BearableSideEffects**. If a side effect and positive word are mentioned we interpret this as meaning that the side effect is present but bearable.
   eg: *So far my joint pain $_{symptom}$ is better$_{positiveWord}$ and my energy and motivation had noticeably improved$_{positiveWord}$*

   $\forall$sentence
   Contains(sentence,Symptoms)
   $\wedge$ Score(sentence,Poswords) > Score(sentence,Negwords)
   $\rightarrow$ ArgumentType(sentence,bearableSideEffects)

6. **UnbearableSideEffects**. If a side effect and a negative word are mentioned we interpret this as meaning that the side effect is present and unbearable.
   eg: *I had several fevers$_{symptom}$ and bone pain$_{symptom}$ making it very difficult$_{negativeWord}$ to get up*

   $\forall$sentence
   Contains(sentence,Symptoms)
   $\wedge$ Score(sentence,Negwords) > Score(sentence,Poswords)
   $\rightarrow$ ArgumentType(sentence,unbearableSideEffectsType1)

7. **UnbearableSideEffectsII**. If a side effect is mentioned in a sentence whose sentiment score is neutral we interpret this as meaning that the side effect is present and unbearable.
   eg: *The constant nightly* $\underline{\textit{hot flashes}_{symptomWord}}$ *and* $\underline{\textit{joint pain}_{symptom}}$ *are* $\underline{\textit{irritating}_{negativeWord}}$ *but yet I'm still* $\underline{\textit{hopeful}_{positiveWord}}$

   $\forall$`sentence`
   `Contains(sentence,Symptoms)`
   $\wedge$ `Score(sentence,Negwords) = Score(sentence,Poswords)`
   $\rightarrow$ `ArgumentType(sentence,unbearableSideEffectsType2)`

8. **PositiveExperience**. The presence of only positive words is interpreted as meaning a positive experience.
   eg: *I felt much* $\underline{\textit{better}_{positiveWord}}$ *on it*

   $\forall$`sentence`
   $\neg$`Contains(sentence,Symptoms)`
   $\wedge$ `Score(sentence,Poswords) > Score(sentence,Negwords)`
   $\rightarrow$ `ArgumentType(sentence,positiveExperience)`

9. **NegativeExperience**. The presence of only negative words is interpreted as meaning a negative experience.
   eg: $\underline{\textit{Terrible}_{negativeWord}}$ $\underline{\textit{terrible}_{negativeWord}}$ *drug*

   $\forall$`sentence`
   $\neg$`Contains(sentence,Symptoms)`
   $\wedge$ `Score(sentence,Negwords) > Score(sentence,Poswords)`
   $\rightarrow$ `ArgumentType(sentence,negativeExperience)`

10. **SuppDrugAvailable**. A sentence containing a symptom and another drug, which is not the drug being reviewed, is taken to mean that the patient is taking a supplementary drug. The predicate `mainDrug(drug)` holds when `drug`, which the drug being reviewed, is not mentioned in the sentence.
    eg: *I have* $\underline{\textit{anxiety}_{symptom}}$ *added* $\underline{\textit{Ativan}_{suppplementaryDrug}}$ *to my drugs...*

    $\forall$`sentence, drug`
    $\neg$`Contains(sentence,Symptoms)` $\wedge$ `Contains(sentence, Drugs)`
    $\wedge$ $\neg$`mainDrug(drug)`
    $\rightarrow$ `ArgumentType(sentence,supplementaryDrugs)`

In this section we have formalised 10 classification rules that are used to extract arguments from medical drug reviews. We show in the next section that our classification rules, albeit simple, yield a reasonable performance. The rules could further be improved by harnessing argument mining techniques and natural language processing.

3 Evaluation of Extracted Arguments

The rules mentioned in the previous section were tested against a set of 570 reviews concerning 4 drugs. In order to validate the performance of each of these rules, the extracted arguments were manually checked by a single human annotator (first author) to see if they had been classed correctly. Each extracted argument was marked as being either *T* (true - the argument was classified correctly), *F* (false - the argument was classified in the opposite class and could in fact be used as a counter argument) and *NA* (irrelevant - argument extracted has no relation with its intended class).

Rule	No. Arguments Extracted	No. T	% T	%F	%NA
PositiveExperience	368	182	49.46	17.12	33.42
NegativeExperience	446	294	65.92	3.81	30.27
NoSideEffectsI	18	18	100	0	0
NoSideEffectsII	31	17	54.84	22.58	22.58
SideEffectsI	142	114	80.28	7.74	11.97
SideEffectsII	61	52	85.25	4.92	9.84
BearableSideEffects	22	11	50	31.82	18.18
UnbearableSideEffectsI	93	81	87.10	4.30	8.60
UnbearableSideEffectsII	320	261	81.56	7.50	10.94
SuppDrugAvailable	180	21	11.67	2.7	85.56

Table 1: Accuracy of all arguments pulled out per classification rule

The results in Table 1 demonstrate that using our classification rules, it is possible to extract relevant arguments regarding treatments. The rules exhibited different precisions (where precision = No. T/No. of Arguments Extracted). For example the rules *NoSideEffectsI* and *PositiveExperience* achieved precisions of 100% and 49.46% respectively. This variability is expected as some of the rules, such as the *PositiveExperience* rule, search for context independent words whereas others search for the occurrence of medical terminology. We also recorded lower precisions when comparing positive sentiment rules to negative ones, eg: *BearableSideEffects* vs. *UnbearableSideEffectsI*. We attributed this to our observation that patients rarely mention a side effect without the intent of complaint.

Alongside these difficulties, we encountered a number of natural language challenges, the majority of which were attributed to the casual nature with which authors wrote their reviews. The difficulties encompassed spelling mistakes, adoption of new terms, abbreviations and general violations of English grammar. Another challenge was the use of non-standard terminology to describe side effects. The quote below highlights this kind of issue.

"...my vision seems to be getting weak."

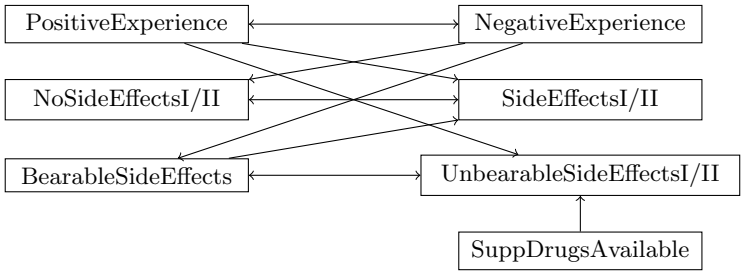

Fig. 1: Argument graph capturing attack relation between the various classification rules.

Discerning a loss of vision from the use of the word *weak* is non-trivial, and is not easily captured using lookup data. Going forward we would seek to improve our classification rules by adopting better natural language processing techniques, and in the case of non standard terminology we could employ techniques such as co-locational data.

4 Evaluation of Argument Graphs

In this section we investigate instantiating an argument graph for each drug review with the arguments extracted from it, we then use Dung's grounded semantics to derive a rating for the drug. We validate these argument-based ratings by correlating them with the numerical ratings given by the authors at the end of their drug reviews.

In order to instantiate the argument graph for a drug review we require a defined set of attack relations for all argument types. In Figure 1 we specify these attack relations based on observations, of a large number of reviews, of how each argument type influences the numerical ratings provided by the user; more specifically we model the competing levels of influence that the argument types have over the rating with respect to one another. As an example *Positive/Negative Experiences* attack all other arguments of opposing polarity to themselves (eg: *NegativeExperience* attacks *NoSideEffectsI/II* and *BearableSideEffects*. This is based on our observation that patients frequently rated in accordance to their overall experience of the drug albeit in the presence/absence of severe/bearable side effects. Other such relationships were observed across the drug reviews and have been represented in our choice of attacks relations.

A consequence of our choice of attack relation is that the grounded extensions of Figure 1 and any of its subgraphs constitute either entirely positive arguments, negative arguments or an empty set. These three possible sets indicate three polarities (positive, negative and neutral) and serve as our argument-based ratings. In order to validate these ratings we correlate the polarity of a drug review to the numerical value provided by the user. In facilitating this correlation the numerical scale was split into three ranges. We assume that a drug review with a

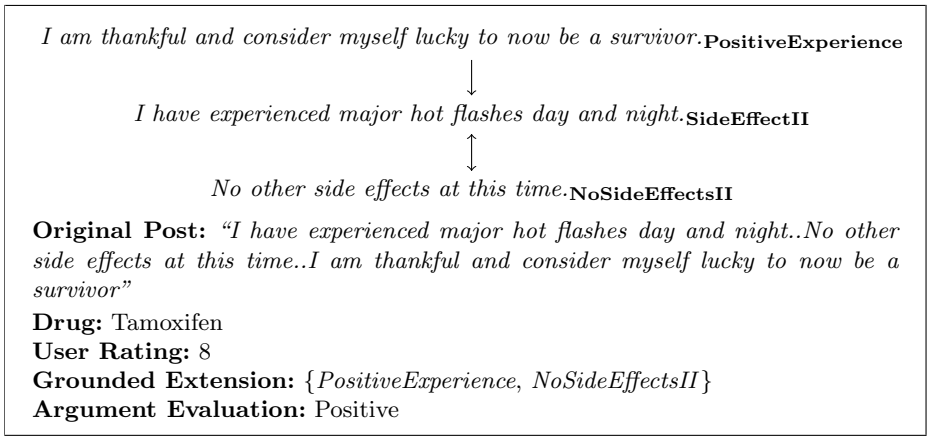

Fig. 2: A review for the drug Tamoxifen. Three arguments were extracted. The grounded extension contains only positive arguments and so the argument-based rating is positive.

rating less than 4 to be a negative rating, a drug review with a rating between 5 and 7 to be neutral and any drug review with a rating greater than 7 to be positive. An example of our system in practice, from argument extraction through to analysis, can be seen in Figure 2.

We ran our experiment using two sets of arguments. In the first set, we used all of the arguments extracted using our classification rules. This was to evaluate the performance of our entire automated process, from argument extraction through to analysis of arguments. In the second set of arguments, we utilised only those extracted arguments which have been annotated as being of type *'T'*. By comparing the correlation matrices for both argument sets we are able to measure the effect of inaccuracies in our classification rules on the argument-based ratings.

Rating	Negative	Neutral	Positive
1-4	0.531	0.443	0.262
5-7	0.198	0.216	0.172
8-10	0.270	0.340	0.566

Table 2: Dung Assesment vs. User Rating using all posts

The results of our experiment in Tables 2 and 3 indicate a positive correlation in the positive and negative classes. It can be seen that there is a notable improvement in correlation in Table 3, given that here we use only validated

Rating	Negative	Neutral	Positive
1-4	0.624	0.417	0.129
5-7	0.206	0.202	0.178
8-10	0.170	0.380	0.693

Table 3: Dung Assesment vs. User Rating using only validated sentences

arguments. The neutral class appears comparatively less correlated with classifications distributed across the ratings scale. What we observed was that reviews whose constituent arguments predominantly shared the same polarity tended to have a numerical score consistent with this polarity. Drug reviews that have neutral numerical ratings often contained predominantly negative or positive arguments causing us to derive a non-neutral argument-based rating. In other cases it was seen that the author would provide positive and negative statements within a single drug review, and whilst the majority of content was homogeneous in its polarity, one statement may have caused the user to rate otherwise.

5 Discussion and Literature Review

In this paper we have presented an argument-based framework for analysing medical drug reviews to be used by patients who are choosing between multiple treatment options. We have shown how simple domain-specific techniques can be used to extract arguments, but this is only so that we have the necessary input for argument-based analysis. Whilst our work is not intended to be a contribution to argument mining, whose motivation is the automated the extraction of argument components, primarily premises and conclusions, from text [3] [11] [6], we acknowledge that techniques from argument mining could be employed to improve our system.

Our work resembles [10] which proposes the use of lookup data in conjunction with argument schemes to mine user generated arguments from online camera reviews. Whilst that paper successfully mines arguments for a specific product, it does not provide an evaluation of arguments mined using any argument solver, whereas evaluating arguments is the primary aim of our paper.

Our approach was to identify a small set of argument types common across all drug reviews and then construct classification rules to extract those argument types. This is in contrast to a manual annotation approach as in [5] which extracted arguments from a set of reviews and put them together in a single argument graph. Our approach enabled us to fully automate our entire system, from extraction through to analysis. It also ensured we had to only construct a single set of attack relations which we imposed on all of our drug reviews.

Going forward we will seek to extend the evaluation of the arguments by making use of the quantity of arguments populated for a given argument class. We will also consider using preference-based frameworks [1], probabilistic frameworks [7] and social abstract argumentation [8] to allow us to model argument

types that are more frequent and yield greater influence over the overall patient ratings. We will also investigate the possibility of learning the attack relations by analysing the numerical rating of a drug review and attempting to construct an argument graph such that we maximise correlations between our argument-analysis rating and the numerical rating.

Acknowledgements

The first author is grateful to the The Royal Free Charity and the EPSRC for funding his PhD studentship. The authors are grateful to the reviewers for their helpful feedback.

References

1. L. Amgoud and S. Vesic. Repairing preference-based argumentation systems. *In Proceedings of International Joint Conference on Artificial Intelligence*, pages 665–670, 2009.
2. J. Cole, C. Watkins, and D. Kleine. Health advice from internet discussion forums: How bad is dangerous? *Journal of Medical Internet Research*, 18(1):e4, Jan 2016.
3. G. P. C.Sardianos, I. Katakis and V. Karkaletsis. Argument extraction from news. *In Proceedings of the 2nd Workshop on Argumentation Mining, Association for Computational Linguistics*, pages 56–66, 2015.
4. P. M. Dung. On the acceptability of arguments and its fundamental role in non-monotonic reasoning, logic programming, and n-person games. *Articial Intelligence*, 77:321–357, 1995.
5. S. Gabbriellini and F. Santini. A micro study on the evolution of arguments in amazon.com's reviews. *In Proceedings of PRIMA 2015: Principles and Practice of Multi-Agent Systems*, pages 284–300, 2015.
6. H. Huangbo and R. Mercer. An automated method to build a corpus of rhetorically-classified sentences in biomedical texts. *In Proceedings of the First Workshop on Argumentation Mining, Association for Computational Linguistics*, pages 19–23, 2014.
7. A. Hunter and M. Thimm. On partial information and contradictions in probabilistic abstract argumentation. *In Proceedings of The 15th International Conference on Principles of Knowledge Representation and Reasoning*, pages 53–62, 2016.
8. J. Leite and J. Martins. Social abstract argumentation. *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, 3:2287–2292, 2011.
9. M. Lippi and P. Torroni. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology*, 16:1–25, 2016.
10. J. Schneider. Semi-automated argumentative analysis of online product reviews. *In Proceedings of COMMA 2012: Computational Models of Arguments*, pages 43–50, 2012.
11. S. Teufel. Argumentative zoning: Information extraction from scientific text. *PhD Thesis, School of Cognitive Science, University of Edinburgh, Edinburgh, UK*, 1999.