Cite: Khoo, C.S.G., Anurup, B., Rasmi, A., Ranjani, T., & Johnkhan, S.B. (2016). Sentence categorization of consumer drug reviews to identify efficacy and side effect information: A preliminary study of psychotropic drugs. In Proceedings of the 7th Asia-Pacific Conference on Library & Information Education & Practice (A-LIEP 2016) (pp. 206-217). Nanjing, China: School of Information Management, Nanjing University.

Sentence Categorization of Consumer Drug Reviews to Identify Efficacy and Side Effect Information: A Preliminary Study of Psychotropic Drugs (Paper ID: 105)

Christopher S.G. Khoo
Wee Kim Wee School of Communication & Information
Nanyang Technological University
chriskhoo@pmail.ntu.edu.sg

Borah Anurup
Wee Kim Wee School of Communication & Information
Nanyang Technological University
anurupborah2001@gmail.com

Ahamed Rasmi
Wee Kim Wee School of Communication & Information
Nanyang Technological University
rasmi_ahd@hotmail.com

Thirunavukkarasu Ranjani
Wee Kim Wee School of Communication & Information
Nanyang Technological University
t.ranjani17@gmail.com

Sathik Basha Johnkhan
Wee Kim Wee School of Communication & Information
Nanyang Technological University
sathik@ntu.edu.sg

ABSTRACT

Background. Consumer drug reviews contain a wealth of information not found on authoritative drug information sites.

Objectives. The paper describes a project to mine consumer drug reviews for drug efficacy and side effect information. The first step in the text mining effort is to categorize sentences in the drug reviews and filter out those containing efficacy and side effect information. The results of an initial study of sentence

categorization on a sample of psychotropic drug reviews (for treatment of mental illnesses) is reported.

Methods. 1000 sample drug reviews were coded by three undergraduate students. 70% of these were used as the training corpus. Automatic sentence categorization models were developed using logistic regression and support vector machine to categorize sentences into effective/positive sentiment, side effect information, and a combined category of negative sentiment or side effect information. The sentence features used in the categorization task were unigrams, bigrams and general features of review length, sentence position, number of matches with entries in a side-effect dictionary, the "side effect" term, and the sentiment score.

Results. The sentence categorization models obtained an accuracy of only 0.62 (F1 measure) in identifying sentences with positive/effective sentiment, and 0.52 in identifying sentences reporting side effect information. Logistic regression analysis found that the sentiment score is a good predictor of positive and negative sentences, and that matches with a side-effect dictionary predicts side effect information as well as positive sentiment. A few counterintuitive predictors were identified as well as interactions between the features.

INTRODUCTION

This paper reports the results of a preliminary study of text mining of consumer drug reviews. Chew and Khoo (2016) had earlier carried out a content analysis of consumer-contributed drug reviews from three websites (WebMD, RateADrug, and PatientsLikeMe), and identified the types of information reported in them that were not generally found in authoritative drug information sites. Types of information found only in consumer drug reviews include drug efficacy, drug resistance, cost of drug, availability of generic versions, comparison with other similar drugs, difficulties in using the drug, and advice on coping with side effects. Chew and Khoo suggested that information on drug efficacy and side effects would be of particular interest to patients and caregivers. Such information, described from the patients' perspective and expressed in layperson's terms, can provide an indication of how fast a drug works and what kind of improvement the patient can expect. It can alert them to potential side effects and drug efficacy in different situations and patient conditions.

Although authoritative drug information sites such as websites of pharmaceutical companies, healthcare institutions, and government health departments do contain comprehensive lists of potential side effects, they are expressed using medical terminology that is difficult for patients to understand and remember. In user drug reviews, drug side effects and their severity are expressed vividly and described in the context of the patient's daily life, together with coping strategies to deal with them.

Doctors and pharmacists can also derive useful information about how particular side effects can affect patients' daily lives, and counsel the patients on appropriate coping strategies. Chew and Khoo found that the drug reviews sometimes reported side effects not stated on authoritative sites. Although such instances were rare, they found a few side effects that had a noticeable number of users reporting them. They suggested that pharmaceutical companies and government regulatory bodies could monitor social media sites for such trends of new side effects surfacing, that were not identified in the clinical trials. This activity has been referred to as pharmacovigilance.

Earlier work on extracting side effect information (also referred to as adverse drug reactions/events) from text have focused on extraction from published medical case reports (e.g., Gurulingappa, Mateen‐Rajpu, & Toldo, 2012) and from database records of adverse drug event reporting systems. Henegar, Bousquet, Lillo-Le Louët, Degoulet and Jaulent (2006) developed an ontology of adverse drug reactions to support the clustering of similar patient conditions to help identify links between drugs and adverse reactions. There is recent interest in the text mining of side-effect information from social media content, but the studies appear to be at an initial phase (e.g., Yang, Jiang, Yang, & Tang, 2012; Yates & Goharian, 2013). Karimi, Metke-Jimenez, Kemp and Wang (2015) constructed a corpus of online postings from the AskaPatient medical forum, with high-quality annotation of concepts of drugs, side effects, symptoms and diseases, to support text mining experiments.

We have embarked on a project to develop a text mining method to extract drug efficacy and side-effect information from consumer drug reviews and to summarize the information. The information extraction and summarization method includes the following text mining steps:

- Sentence categorization—to identify and filter out sentences in the drug reviews containing positive or negative drug efficacy information and side effect information.
- Information extraction—to extract specific side effects and efficacy information from the identified sentences
- Information summarization—to summarize the extracted information in a tabular, textual or visual form.

This paper describes an initial study of the first step—sentence categorization on a sample of drug reviews of psychotropic drugs (for treatment of mental illnesses). The list of psychotropic drugs was constructed from several online drug information sources. Consumer reviews of these psychotropic drugs were filtered out from a corpus of drug reviews from WebMD (http://www.webmd.com/drugs/index-drugs.aspx), constructed by Dr Pauline Ng of the Genome Institute of Singapore. A random sample of 1000 reviews were taken from this subset of psychotropic drug reviews, and segmented into sentences using a program developed in-house. The sentences were coded by three undergraduate students, and a random

sample of 700 reviews (70%) were used as the training set and the remaining reviews as the test set.

It is well-known that online consumer reviews contain consumers' opinion and sentiment towards the product or service being reviewed. Similarly in drug reviews, reports of drug efficacy and side effects are often accompanied by expressions of positive sentiment (the drug is effective or has positive effects) or negative sentiment (the drug is not effective). Side effects are usually negative or undesirable, but users sometimes report unintended positive effects. Thus the study includes a sentiment analysis component.

METHOD

Coding of Drug Reviews

Three undergraduate students from the School of Communication & Information at Nanyang Technological University, Singapore, were recruited to code the sentences in the 1000 sample drug reviews into five categories:

- Patient history of disease or disease symptoms
- Effective or positive sentiment
- Negative sentiment: does not mention a specific side effect, but expresses negative sentiment on some aspect of the drug
- Side effect information: one or more specific side effects are mentioned
- Not on this drug: the sentence mentions or comments on another drug, possibly a drug that the patient previously took.

A sentence can be coded with multiple categories. For example, it may contain both positive and negative information. An exception is when a sentence is coded with Not on this drug, in which case no other category can be assigned to this sentence. After coding, a combined negative sentiment or side effect category was derived, as the side effect category is deemed to be generally negative as well.

Intercoder reliability measures for the three sets of coding are given in Table 1. Average percent agreement is easier to understand, but it does not compensate for chance agreements and coder bias. Krippendorff's Alpha (Krippendorff, 2004a & 2004b; Hayes & Krippendorff, 2007) is the generally preferred measure for content analysis coding (Artstein & Poesio, 2008). A value greater than 0.8 is considered good reliability. A value of 0.67 to 0.8 can be considered acceptable for tentative conclusions (Krippendorff, 2004).

Table 1. Intercoder reliability measures		
Category	**Krippendorff's Alpha**	**Average Percent Agreement**
Patient history	0.42	92%
Effective/positive	0.76	91%
Negative	0.50	93%
Side effect	0.67	91%
Combined negative or side effect	0.67	88%
Not on this drug	0.44	97%

The intercoder reliability values of 0.76 for effective/positive sentiment category and 0.67 for side-effect information is acceptable for a preliminary study. Some coding issues were identified:

- Side effect information category: Some coders included sentences that mentioned that there were side effects, without specifying the particular side effect. Positive side effects were also sometimes included. Coders were uncertain whether to include withdrawal symptoms in this category.
- Patient history of disease or disease symptoms: Coders were uncertain whether the name of the disease (i.e. the diagnosis) and previous drugs used were to be included in this category.
- Not on this drug: Coders were uncertain whether other drugs taken in combination with the reviewed drug was included in this category, and whether drug interaction effects were in this category.

The sentence categorization experiments were carried out using the following categories:

- Effective or positive sentiment
- Side effect information
- Combined negative sentiment or side effect information.

From the three sets of manual coding, sentence categories selected by at least 2 coders were taken as the gold standard.

Feature Extraction and Selection

The sentences were lemmatized (converted to root form) using the Stanford core NLP parser (Manning, Surdeanu, Bauer, Finkel, Bethard, & McClosky, 2014), and unigrams (single words) and bigrams (two-word terms) were extracted from the sentences, and used as features to represent the sentences. The sentence frequency for each unigram and bigram was compiled, and the most frequent 10 unigrams were reviewed to form the stoplist. Unigrams and bigrams with sentence frequency less than 3 were also dropped. The remaining 3722 unigrams/bigrams were used as features to construct the sentence vectors to use in model

building. Two kinds of term weighting were investigated: binary weighting (whether the term occurs in the sentence), and binary/(log2 sentence frequency).

As the annotated corpus is small, we investigated the following more general features:

- Review length: number of sentences
- Sentence position: sentence number
- Normalized sentence position: sentence number divided by review length
- Number of side effects mentioned: number of matches with entries in a side-effect dictionary that we had earlier compiled
- Presence of "side effect" term: whether the sentence contains the term "side effect" and its variations (e.g., "side-effect" and "sideeffect")
- Sentiment score: count of positive words minus count of negative words, using the WKWSCI Sentiment Lexicon (Khoo, Johnkhan & Na, 2015; Khoo & Johnkhan, under review) with negation handling

The side-effect dictionary is a list of side effects and variations in expression that we had compiled from online sources, including the U.S. Food and Drug Administration (http://www.fda.gov/), U.S. National Institute of Health (http://www.nih.gov/), WebMD (http://www.webmd.com/), Physician Desk Reference Health (http://www.pdrhealth.com/), and the U.K. Electronic Medicines Compendium (eMC) (http://www.medicines.org.uk).

The difference between sentence position and normalized sentence position can be seen in the way 1-sentence reviews and short reviews are handled, in comparison with long reviews. Sentence position considers the single sentence in 1-sentence reviews as equivalent to the first sentence of longer reviews: sentence position = 1. On the other hand, the normalized sentence position (calculated as m/n, where m is the sentence position and n is the review length) considers the 1-sentence review as equivalent to the last sentence of a longer review: in either case m=n and the normalized sentence position = 1.

EXPERIMENTS

Investigating General Features Using Logistic Regression

Stepwise logistic regression was applied to the training corpus to develop classifiers for the three sentence categories: effective/positive sentiment, side effect information and combined negative sentiment/side effect information. The classifier models also indicate which combination of the general features are significant predictors of the categories. The resulting logistic regression models are given in Table 2.

Table 2. Logistic regression models for the three target variables

	B	Wald	df	Sig.
Target=Effective/positive sentiment				
ReviewLength	-.066	65.947	1	.000
NumberSideEffects	.087	28.872	1	.000
SentimentScore	.942	108.410	1	.000
SentimentScore*NormalizedSentencePosition	-.579	23.412	1	.000
Constant	-.797	80.617	1	.000
Target=Side effect information				
ReviewLength	-.022	6.214	1	.013
NumberSideEffects	.275	188.351	1	.000
SideEffectTerm	1.036	32.002	1	.000
SentimentScore*NormalizedSentencePosition	-.377	33.484	1	.000
SentimentScore*NumberSideEffects	.036	36.151	1	.000
Constant	-1.339	40.578	1	.000
Target=Combined negative sentiment or side effect information				
ReviewLength	-.028	12.411	1	.000
NormalizedSentencePosition	.421	6.932	1	.008
NumberSideEffects	.161	75.156	1	.000
NumberSideEffects*SideEffectTerm	.228	33.340	1	.000
SentimentScore	-.135	3.005	1	.083
SentimentScore*NormalizedSentencePosition	-.308	9.220	1	.002
SentimentScore*NumberSideEffects	.030	23.844	1	.000
Constant	-1.760	135.752	1	.000

Effective/Positive Sentiment Category

Table 2 indicates that the equation for estimating the probability that a sentence is categorized as effective/positive sentiment is as follows:

Log odds (sentence category is effective/positive sentiment) =

– 0.80 + 0.94*SentimentScore – 0.07*ReviewLength + 0.09*NumberSideEffects –

0.58*SentimentScore*NormalizedSentencePosition

The "log odds" value can easily be converted to a probability value.

Table 2 indicates, not surprisingly, that the sentiment score is the most significant variable. Furthermore, longer reviews are less likely to be positive.

Counterintuitively, a higher number of side effect matches is associated with a higher probability of positive sentiment. Here are some examples of sentences with mentions of several side effects, but also contains some positive sentiment:

- headaches post menopausal bleeding anxiety feelings of rage sleeplessness decreased appetite nightmares -- however decreased thoughts of suicidal ideation.
- It helps me sleep but wake up very tired lasts all day makes my pain from Fibromyalia worse dry mouth bad taste in my mouth and now after 10 days I itch .

- It works for my depression but I've been aware of eye problems and other common side effects such as sexual ringing in the ears unusual dreams dry mouth and constipation.
- But to my surprise this medication has worked wonders on me !!! I have anger issues linked to depression/anxiety and I no longer go into fits of rage or crying spells or horrible mood swings.

The first three examples balance the positive effect of the drug with a list of side effects. The last example gives a list of symptoms that had been negated by the drug.

There is also an interaction between sentiment score and normalized sentence position: a higher sentiment score in the last sentence of a review (or in a 1-sentence review) indicates less positive sentiment. This may be an adjustment (a "discount") to the effect of the sentiment score for sentences towards the end of a review. Long reviews tend to have more narrative, advice and sarcastic texts towards the end of the review, that are harder to categorize, as in the following examples:

- Adderall is 4 amphetamine salts in one tablet so it's like the inventors of this drug just said to themselves "well one of these stimulants has to hit the right spot" and then they marketed it with the clever name ADD ERALL. [Sentence no. 14 of 17 sentences]
- Please find other FDA approved tried and true alternatives to this terrible substance (under the care of a good quality physician). [Sentence no. 12 of 15 sentences]

The classifier was applied to the test corpus to evaluate its predictive accuracy. A threshold probability of >0.25 was used to categorize a sentence as belonging to the effective/positive category, on the basis that 25% of the sentences in the training corpus had been manually assigned to this category. The classifier obtained a recall of 73% and precision of 42%, yielding an F1 score of 0.53.

Side Effect Information Category

Table 2 indicates, not surprisingly, that number of side effects is the most significant variable. Sentences containing the term "side effect" also tended to be assigned to this category.

Long reviews tended to have a lower percentage of sentences describing side effects. Table 3 indicates that 1-sentence reviews in the training sample describe side effects 23.4% of the time. For reviews of length 2 to 12 sentences, the percentage of sentences describing side effects range from 14% to 25%. For reviews longer than 12 sentences, the percentage of sentences describing side effects drop to 10%.

The sentiment score is also a useful predictor, but only in combination with the normalized sentence position and the number of side effects. Positive sentences towards the end of a review (or in 1-sentence reviews) were less likely to contain side effect information. Counterintuitively, a higher sentiment score coupled with higher number of side effects

increases the probability of containing side effect information. The following examples suggest this is because of sentences that report drug efficacy balanced with side effects:

- The medication was effective however the side effects is great Not able to sleep not able to eat then when the appetite returns hve a craving to eat 24*7 unusual dreams that seem to be real
- This drug is very effective easy to use and overall we are satisfied but it is true what they say about weight gain in teens.

Applying the classifier to the test corpus using a threshold of >0.17 obtained a recall of 65% and precision of 30%, yielding an F1 score of 0.41. It is clearly difficult to identify sentences containing side effect information.

Table 3. Average percentage of sentences describing side effects for different review lengths

Review Length	Avg % of sentences containing side effect information	No. of reviews (N)
1	0.234	145
2	0.176	108
3	0.142	82
4	0.187	91
5	0.178	68
6	0.192	42
7	0.206	38
8	0.161	21
9	0.182	25
10	0.245	11
11	0.219	15
12	0.204	9
13	0.062	9
14	0.127	8
15	0.058	7
16 to 29	0.113	21
Total		700

Combined Negative Sentiment or Side Effect Information

As this category is a mixed category, the logistic regression model in Table 2 is more complicated, with more features. All the general features were found to be significant predictors. The most important feature was again the number of side effects, especially if they were accompanied by the term “side effect”. A positive sentiment score reduces the probability of the sentence being negative/side effect, especially for sentences towards the end of a review (or in 1-sentence reviews).

Interestingly, both the review length and the normalized sentence position were significant. The negative coefficient for review length indicates that longer reviews tended to contain fewer negative sentences. But sentences towards the end of a review (or 1-sentence reviews) tended to be negative.

Applying the classifier to the test corpus using a threshold of >0.24 obtained a recall of 63% and precision of 37%, yielding an F1 score of 0.47.

Sentence Categorization Experiments Using Support Vector Machine

As the general features were not sufficient to give good categorization accuracy, we carried out a text categorization experiment using the general features together with unigrams and bigrams as features. The support vector machine (SVM), using a linear kernel, was applied to the training set to develop classifiers. The sentence categorization results when the classifiers were applied to the test set are given in Table 4.

Looking at the F1 scores in Table 4, the SVM models using unigram/bigram features were clearly better than the models using only general features for the effective/positive and side effect categories. Adding the general features to the unigram/bigram features also did not noticeably improve the accuracy. For the combined negative or side effect category, the model using unigram/bigram features had about the same F1 score as using the general features, and combining them did improve the F1 score from 0.45 to 0.49. We credit this to the use of the WKWSCI Sentiment Lexicon to identify negative opinions.

Table 4. Sentence categorization results using support vector machines

	General features only			**General features + unigrams + bigrams**			**Unigrams + bigrams**		
Target category	**Recall**	**Prec.**	**F1**	**Recall**	**Prec.**	**F1**	**Recall**	**Prec.**	**F1**
Effective or positive	0.70	0.42	**0.52**	0.60	0.63	**0.62**	0.59	0.62	**0.60**
Side effect information	0.64	0.30	**0.41**	0.61	0.45	**0.52**	0.61	0.45	**0.52**
Negative or side effect	0.61	0.37	**0.46**	0.53	0.45	**0.49**	0.50	0.42	**0.45**

CONCLUSION

It is not an easy task to extract efficacy and side effect information from consumer drug reviews. Our sentence categorization models obtained an accuracy of only 0.62 (F1 measure) in identifying sentences with positive/effective sentiment, and 0.52 in identifying sentences reporting side effect information, using a training corpus of 700 reviews of psychotropic drugs.

The SVM and logistic regression models using general review and sentence features (i.e. review length, sentence position, number of side effects mentioned, the “side effect” term, and sentiment score) obtained generally weaker results than SVM models using

unigram/bigram features. However, the logistic regression analysis with the general features identified interesting characteristics of the sentences:

- A higher (i.e. more positive) sentiment score predicts positive sentences that indicate the drug is effective. Lower (more negative) sentiment score predicts negative sentences. However, the sentiment score interacts with the normalized sentence position and the number of side effects in unexplained ways.
- The number of matches with our side effect dictionary naturally predicts a sentence with side effect information. However, it sometimes also suggests positive effects and drug efficacy.
- Longer reviews negatively predicts both positive and negative sentences! Perhaps the longer reviews contain more patient history information and coping advice.
- The normalized sentence position is a predictor of negative sentences: Negative sentences tend to occur in 1-sentence reviews or towards the end of longer reviews.

We are currently carrying out qualitative content analysis to understand the reasons for these results. As it is difficult to develop a high-quality annotated corpus to carry out text categorization and information extraction experiments, we are investigating methods to bootstrap a classification model with only a small amount of manual coding, and the automatic construction of a training corpus.

ACKNOWLEDGEMENT

We acknowledge with grateful thanks the use of a consumer drug review corpus constructed by Dr Pauline Ng of the Genome Institute of Singapore.

REFERENCES

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4), 555-596.

Chew, S.H., & Khoo, C.S.G. (2016). Comparison of drug information on consumer drug review sites versus authoritative health information websites. *Journal of the American Society for Information Science and Technology*, 67(2), 333-349.

Gurulingappa, H., Mateen-Rajpu, A., & Toldo, L. (2012). Extraction of potential adverse drug events from medical case reports. *Journal of Biomedical Semantics*, 3(1), 1.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1, 77-89.

Henegar, C., Bousquet, C., Lillo-Le Louët, A., Degoulet, P., & Jaulent, M. C. (2006). Building an ontology of adverse drug reactions for automated signal generation in pharmacovigilance. *Computers in Biology and Medicine*, 36(7), 748-767.

Karimi, S., Metke-Jimenez, A., Kemp, M., & Wang, C. (2015). CADEC: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, 55, 73-81.

Khoo, C.S.G., & Johnkhan, S.B. (under review). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons.

Khoo, C.S.G., Johnkhan, Sathik B., & Na, J.C. (2015). Evaluation of a general-purpose sentiment lexicon on a product review corpus. In R.B. Allen, J. Hunter, & M.L. Zeng (Eds.), Digital libraries: Providing quality information: 17th International Conference on Asia-Pacific Digital Libraries, ICADL2015: Proceedings (LNCS 9469, pp. 82–93). Berlin: Springer.

Krippendorff, K. (2004a). Content analysis: An introduction to its methodology. Thousand Oaks, CA: Sage.

Krippendorff, K. (2004b) Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*. 30(3), 411-433.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 55-60).

Yang, C. C., Jiang, L., Yang, H., & Tang, X. (2012, August). Detecting signals of adverse drug reactions from health consumer contributed content in social media. In Proceedings of ACM SIGKDD Workshop on Health Informatics. New York: ACM.

Yates, A., & Goharian, N. (2013, March). ADRTrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites. In European Conference on Information Retrieval (pp. 816-819). Berlin: Springer.