Classification of comment helpfulness to improve knowledge sharing among medical practitioners.

Pierre André Ménard
Computer research institute of Montréal
pierre-andre.menard@crim.ca

Caroline Barrière
Computer research institute of Montréal
caroline.barriere@crim.ca

Abstract

Clinical research article summaries called infoPOEMs (Patient-Oriented Evidence that Matters) are emailed by the Canadian Medical Association to family physicians who read them and answer the online Information Assessment Method (IAM) questionnaire which a free form textual opinion fields to comment on the value or content of the infoPOEM. This article presents results of a relevance evaluation study applied on these comments to automatically determine their helpfulness and consequently the interest of sharing them among the medical community. A dataset of 3,470 manually annotated comments provides a gold standard, containing structural, syntactic, and semantic features taken from the Unified Medical Language System and IAM questionnaire. Applied machine learning algorithms show a global f-measure improvement of 9.1% when compared to a binary occurrence bag-of-word baseline.

1 Introduction

The task of opinion mining has gained importance in the last years with our world being increasingly made of posted information with crowds commenting on such information. Such increased volume of crowd comments has led to text analysis research aiming at understanding and clustering the opinions found in those comments (e.g. see recent articles (Mukherjee and Liu, 2012; Turney, 2002; Chen and Zimbra, 2010)) and to help manage interactions within the online community (Huh et al., 2013).

An even more recent task is not so much on understanding comments content, but rather on evaluating comments value, impact or helpfulness for the community reading them. Most research addressing this task, as shown in the Related work section, uses comments on product information on Amazon. But the idea of evaluating comments helpfulness can be extended to other contexts such as community learning or sharing of knowledge. In the present research, we look particularly at the community of medical practitioners in the context of reading and commenting on scientific article summaries which are called infoPOEMs® (Patient-Oriented Evidence that Matters). Within this medical community, comments about an infoPOEM made by one practitioner could be useful to other practitioners regardless of the opinion expressed. The helpfulness dimension is not necessarily correlated with the opinion dimension with typical values being positive, negative or neutral. For example, the comment ”good article” certainly has a different helpfulness value than the comment ”this is a very good article since it shows for the first time that drug X can be useful in disease Y”, even though both comments are positive.

The automatic identification of helpfulness becomes the subject of our research. Practitioners are not interested in reading all comments, only the valuable or ”helpful” ones, and an automatic identification of helpfulness would provide a more efficient way for knowledge sharing among them.

2 Related Work

Assessing the helpfulness of comments made about recreational or informational items has been explored recently mainly for online products or movie reviews. Many studies look at Amazon data, which is perfectly suited for this task since it provides training data readily available. Besides the research work on Amazon data, we also mention one work on peer review in an educational context. We provide pretty extensive details on the features selected and results obtained for the different work to allow us to compare our feature

Proceedings of NAACL-HLT 2016, pages 72–81,
San Diego, California, June 12-17, 2016. ©2016 Association for Computational Linguistics

sets and our results to the ones mentioned here.

Within the Amazon studies, the most cited approach by Kim and Pantel (2006) uses machine learning algorithms with text-based features to rank the usefulness of products reviews from the Amazon.com website. They use a dataset of 25,841 reviews from 1,802 products (mp3 players and digital cameras). Their gold standard ranking is based on user responses, provided on the site, to the question ”Was this review helpful to you?”. Their features are divided into five sets: structural, lexical, syntactic, semantic and metadata. Structural features comprise total token number, number of sentences, average sentence length, percentage of question sentences, number of exclamation sentences, bold and line break html tags. Lexical features comprise tf-idf (Term Frequency-Inverse Document Frequency) of unigrams and bigrams. Syntactic features included percentage of open-class tokens, verb tokens, first person verbs, adjective and adverb. Semantic features comprise occurrences of product features in the review, as well as occurrences of positive and negative sentiment words. Metadata features comprise number of stars rating, difference between given star rating and average star rating of the product. Using these comments' derived features, they use a SVM-RBF algorithm to evaluate features' correlation. Their best result used only three features (comment's length, unigrams and star rating) providing a Spearman rank correlation of 0.66.

Ngo-Ye and Sinha (2012) also looks at Amazon.com reviews, using 2,718 reviews of 11 books. Rather than expanding the set of features, as with Kim and Pantel (2006), they limit themselves to the traditional bag-of-words approach. Their contribution is on dimensionality reduction using the regressional ReliefF algorithm in comparison with LSA, correlation feature selection (CFS) and two other dimension reduction methods. Using both binary occurrences of words and real frequency occurrences as features, they conclude that the use of regressional ReliefF dimension reduction algorithm outperform basic bag-of-word, LSA and CFS on every count.

Zhang and Tran (2008) suggests an information entropy-based bag-of-word model to predict the helpfulness of reviews. As training data, they use 9,955 gps and mp3 player reviews from Amazon. Contrarily to Kim and Pantel (2006) who attempted correlation with a gold standard ranking, they transform the problem into a binary classification problem using a consumer vote ratio threshold of >60% to consider a review as helpful. They compare their entropy-based method with three classifiers: Naive Bayes, Decision Tree and sequential minimal optimization (Platt, 1998). The resulting performances (77.2% for helpful and 77.5% for non-helpful) for their approach beat Naive Bayes (h:76.2% and n-h:75.2%) and Decision Tree (h:72.3% and n-h:75.3%) but of the same rank as an occurrence-base bag-of-word using the SMO classifier (h:76.1% and n-h:78.0%) when considering both value of the output class.

Other research also take place in other fields like educational peer-review systems. This is the case with Xiong and Litman (2011) who used a feature-based machine learning approach on peer-review assessments from an introductory collegial history class to evaluate their usefulness. They collected 267 comments made on 16 papers which evaluated the quality of the work (facts, clarity, argument structure and so on). While using the previously published features (Kim and Pantel, 2006) as a baseline, they introduced new features like problem localization (”Page 2 says ...”), new lexicon categories (modal verb, negation, positive and negative words, ...) and cognitive-science constructs (praise, problem, summary, solution, ...). The baseline using structural, unigrams and metadata features offered a 0.62 Pearson correlation (0.67 with new features). The context of this research is the nearest to ours as it targets the usefulness of comments for educational purposes.

3 Applicative context

The Canadian Medical Association (CMA) delivers by email clinical research article summaries called infoPOEMs (Patient-Oriented Evidence that Matters) to family physicians around the country. To transform the reading of infoPOEMs into an actual learning experience, research in education states the importance of having the reader (learner) reflect on the value of his reading by answering questions. While questions on the content only test short-term memory, questions on the value of the information for clinical practice can stimulate reflective learning. The impact of such practice and its validation have been researched in depth (Grad et al., 2006; Grad et al., 2008; Pluye et al., 2010a; Pluye et al., 2010b).

As part of their mandatory continuing education program, physicians can answer an online questionnaire called Information Assessment Method or IAM (Grad et al., 2011). It contains many questions to gauge the impact of the infoPOEM's content on the physicians knowledge and practice: ”Is your practice changed and improved?”, ”Are you motivated to learn more?”, ”Are you dissatisfied?”, ”Is this summary relevant for at least one of your patients?”. In addition to the predefined questions, physicians can add comments about their reading experience targeting the quality of the overall infoPOEM information, the research, the methodology, and so on. The examples below illustrate how physicians' comments can fluctuate in length, content and targeted issues.

1. Content of drops not specified
2. Why was this study done when we have prev information regarding pot harms done with acute lowering of BP post stroke?
3. Good to hear this as CRP is a rather non-specific marker
4. very interesting
5. Cost of each Rx regime?
6. administrative physician

Comments can be related to missing information (1, 5), generic appreciation (4), critical disagreement (2), agreement with support (3), contextual information about inapplicability of the information (6), etc.

4 Methodology

Our research takes a similar approach as Kim and Pantel (2006) and Xiong and Litman (2011) on feature extraction and machine-learning, while looking at a closed system without clear ”wisdom-of-the-crowd” indicators. We evaluate the impact of features based on textual analysis of the comment itself, but also features based on a comparison between the infoPOEM and the comment, as well as features relying on external domain-specific resources. Our methodology consists of (1) circumscribing the data and developing a gold standard, (2) defining a set of features that will best describe the data to be categorized, (3) experiment with machine learning approaches for categorization and (4) perform an evaluation using the gold standard.

4.1 Dataset and gold standard

The gold standard was annotated by three medical students with different experience levels. They were asked to read anonymous comments submitted by physicians and indicate if they found them valuable for their knowledge or practice. [1]

Each annotator was provided with a list of anonymous comments and their associated infoPOEM for reference. They could access, if needed, the full text of the infoPOEM if the comment was not clear to them. A preliminary annotation phase was done with 300 randomly selected comments to be annotated by the three annotators (100 each). This phase provided a better understanding of the problem to validate the annotation schema used for the main annotation task. The classification schema included three choices to annotate the helpfulness of a comment: ”valuable”, ”non-valuable” or ”I don't know”. The annotators were asked to consider each comment independently and not let the reading of previous comments influence their choice.

The main annotation task was based on two batches of comments. A first one, relatively small, contained 250 comments and was given to all three reviewers and allowed us to calculate an inter-annotator agreement. A larger set of comments was split in three parts to have each comment annotated by a single reviewer. This provided a total of 3,470 comments associated with 327 randomly picked infoPOEMs. Of these comments, 1,586 (45.6%) were deemed valuable and 1,884 (54.3%) non-valuable. A dozen comments were tagged ”I don't know” and removed from the dataset.

The 300 comments from the preliminary annotation step joined with the 250 comments for the inter-annotator agreement were used as the development dataset (550 unique comments) to define, develop, test and refine features presented in the next section but were not used in the dataset for the final evaluation. The other set of 3,470 comments was used as the evaluation dataset for performance assessment.

The size of the manually annotated dataset compares advantageously to the 1000 annotated comments of Ghose and Ipeirotis (2007) and the 267 of Xiong and Litman (2011). Using the first 250 comments annotated by the three annota-

[1] The anonymous comments were provided by the Canadian Medical Association under a non-disclosure agreement for research purposes only.

tors, an inter-annotator agreement of 0.4846 was computed using the Fleiss' Kappa method for multiple annotators with all three classes (valuable / non-valuable / i don't know). The inter-annotator agreement was recalculated using only the 247 comments with only the two main classes (valuable/non-valuable) which provided a score of 0.5004. The remaining data shows a stronger agreement on valuable comments than on non-valuable ones. The level of agreement calculated on this dataset is considered moderate according to Landis and Koch (1977) when compared to pure chance agreement and is of the same order as in Xiong and Litman (2011). Using each annotator as the gold standard versus others, the f-measures were 0.806 between annotators 1 and 2, 0.783 between 1 and 3 and 0.792 between 2 and 3.

The reason behind the average ratings for inter-annotator agreement score can be explained by one or many of the following points: coding instructions were interpreted differently by each annotator, coding decision is based on factors which are not present in textual data (like relevant prior knowledge, expertise domain or interest, personal taste or bias and so on), decision factors were present in the text but not correctly understood by the readers, etc. While it is difficult to provide a clear and proven diagnosis of the reason behind these scores, lower scores usually increase the difficulty to develop prediction systems. As such, the average agreement provides a contextualisation of potential performance for this task; a near-perfect classification of comments is not the goal as it would overfit the three annotator's classification.

4.2 Feature definition

The purpose of defining features is to capture as well as possible the characteristics of comments which would be representative of their helpfulness character. Inspired by previous research, we define a set of base features, focusing on standard text analysis techniques. But we apply these techniques not only to the comment's content itself, but also in a comparative setting looking at similarities between an infoPOEM and its comments. We present these base features first. Second, we look at metadata features from the infoPOEM itself. Third, we use the actual IAM questionnaire as a source of features. Fourth, inspired by our specific problem being in the medical domain, we define a set of features using a medical resource, the UMLS (Unified Medical Language System). The feature extraction process was developed using GATE (Cunningham et al., 2011) with part-of-speech TreeTagger (Schmid, 1994) tool.

4.2.1 Base

The base set includes all features extracted using natural language processing techniques. It includes features and their representations used in previous researches like Kim and Pantel (2006; Xiong and Litman (2011) as well as new ones introduced in this article. They can be regrouped in the structural, syntactic and semantic subsets.

Structural Structural features target statistical properties of tokens contained in the comments. The total number of each one was added as separate features. Two features were also added for tokens: the standard deviation and a three-value discretization of the standard deviation to account for the length being within range of the average (*avg*) number of tokens of all comments, above (*high*) or under (*low*) it, using $\pm 1\sigma$ as the threshold.

Syntactic Following a part-of-speech tagging (attributing a syntactic role to each word), the number of stop words and content words were added as features, which summed up to the number of tokens from the structural feature. The standard variation and its discretization (as seen previously) were also added. The first and second person pronouns (ex: I, we, us, etc) were added as total count and binary occurrence (true if any occurrence are observed, false if none) features to the dataset to identify author related comments like accounts of personnal experiences, thoughts, preferences or opinions.

Then for each type of content words (verb, adverb, noun, adjective) found both in the comment and the corresponding infoPOEM, we added four similarity-based features. They were the total count of similar occurrences, the binary occurrence, the ratio between the total count and the total number of content words and finally the ratio between the total count and the total number of words.

Semantic To identify comments with strong opinions or impressions, we use specific verbs (e.g. admit, enjoy, deem, endorse, decline, concern, advise, ...) and match the infinitive form of these verbs in the comments following a part-of-speech tagging step. Negative indicators (not, never, neither, nor, can't, don't, etc) are also annotated to target potentially critical comments. As the comments were on infoPOEMs within a scientific discipline, terminology related to the scientific method (observation, qualitative, inference, ...), the statistical domain (population, marginal variable, match sample, ...) and to measurement (unit, cm, m, mg, ug, kg, ml, ...) were added separately as features. Finally, the five standard section's labels (title, clinical question, bottom line, study design, synopsis) from the infoPOEM were added as keywords to detect if a text was commenting on the specific section of the infoPOEM.

The number of instances and the binary occurrence for each of these semantic concepts (opinion verbs, domain terminology, negative indicators and localisation indicators) were added as features.

4.2.2 Metadata

To each infoPOEM is associated a code called the level of evidence (LOE). This code describes the type of research protocol used in therapy, diagnosis or prognosis research using one letter and one number (1a, 1b, 1c, ..., 2a, 2b, ...). A minus sign can be added at the end of the code to denote researches that cannot provide conclusive answers in cases where the confidence interval is too large or the heterogeneity of the population's sample used is problematic. We use this code and split it in 3 parts to provide 3 features: the type (first character, from 1 to 5), the subtype (second character, from A to C) and the presence of the minus indicator.

4.2.3 IAM

Each question from the IAM questionnaire was added as a feature. Most of the questions asked for a logical yes/no answer. A few questions accepted either yes, no or "possibly" as an answer. Only one question pertaining to the relevance of the information regarding the physician's patients, asked for an answer using three levels: "totally relevant", "partially relevant" or "not relevant". The possibility to answer some specific questions was also dependant on the answer on previous questions; i.e. questions #3 and #4 were only available if the totally or partially relevance was chosen at question #2. Regardless of this factor, all questions were added as individual and stand-alone features in the dataset.

4.2.4 UMLS

Unlike the work with Amazon data which relies on official product feature sources to find vocabulary representative of different products, we do not have access to such sources in this study. Instead, we extracted single words and multiword expressions from the Unified Medical Language System, a large medical ontology hosted at the National Library of Medicine (`http://umlsks.nlm.nih.gov/`) to analyse the domain specific nature of the reviews and infoPOEMs. The relevant part of this resource splits biomedical and related concepts into 13 groups and 94 types using themes like genes and molecular sequences, anatomy, living beings, physiology, procedures, disorders, organizations and so on, with each type related to one group.

For each type and group, the number of occurrences, the binary occurrence and the similarity occurrences were added as features. The similarity occurrence indicates how many expressions found in a comment were also found in the infoPOEM related to that comment. This type of feature was added to verify if an author was talking about domain-specific concepts from the infoPOEM. Because of the relation between groups and types, each matching expression was both represented with a type feature and its corresponding group feature. In addition, the global binary and total occurrence of UMLS expressions were added as two features to logically regroup all UMLS type and group features. Therefore, if a word was tagged as being part of 4 types and 3 groups, the global binary occurrence would be 1 and the global number would be 7.

5 Performance evaluation

5.1 Baseline

Two baselines were created using the bag-of-word model applied to the whole set of annotated comments. The preprocessing included a tokenizer, an English stop word filter and the Snowball English stemmer, using each stemmed token as a feature in the dataset. The bag-of-word baselines have been extracted and tested using the RapidMiner tool (Mierswa et al., 2006).

The first baseline follows the best replicable results from Kim and Pantel (2006) using the length in token of the comment and a unigram bag-of-words. The stemmed tokens were then weighted using the tf-idf measure. The resulting dataset was

Table 1: Weighted f-measure results for algorithms on each dataset.

	B	B+I	B+U	B+I+U	150R
BayesNet	0.651	**0.693**	0.673	0.692	0.651
Voted Percept.	0.659	0.692	0.686	0.704	**0.713**
JRip	0.663	0.686	0.671	0.679	**0.688**
LMT	0.660	0.700	0.694	0.700	**0.708**

(B): Base, (I): IAM questionnaire, (U): UMLS,
(150R): B+I+U with selection of 150 features with Relief-F

processed with the SVM-RBF algorithm which provided a f-measure score of 63.6%.

The second baseline is based on the conclusion of Zhang and Tran (2008), which presented a method providing a weighted score equivalent to the SMO algorithm applied to a binary occurrences bag-of-words. The SMO (sequential minimal optimization algorithm for training a support vector classier) and binary bag-of-words method performed on our comment corpus yielded a 62.2% f-measure which is significantly lower than the 77.1% f-measure averaged from the helpful and not helpful classes using their product review dataset.

5.2 Helpfulness prediction

As the helpfulness evaluation was based on few annotators instead of large population like on Amazon, classification algorithms were used to predict the correct value instead of a rank correlation method. We used the ten-fold cross-validation to provide the recall, precision and f-measure estimation for each one. The feature sets were then combined with one another to verify which group gave the best results. Evaluation of machine-learning algorithms has been made using the Weka toolset (Hall et al., 2009).

To be able to test the relative strength of the feature sets from section 4.2, four datasets were created using the following feature sets: base (B), base and IAM questionnaire (B+I), base and UMLS (B+U) and all three sets together (B+I+U). The three metadata features they were included in the base feature set (B). Finally, as the UMLS set contains a large number of features, a fifth dataset (B+I+U-150R) was created using a smaller subset of features which were selected following Ngo-Ye and Sinha (2012) study, using the Relief-F algorithm to select the 150 top features, excluding the output class.

For each of the five datasets, a single algorithm from the four main families was tested: BayesNet (Friedman et al., 1997) (baysian), Voted Perceptron (Freund and Schapire, 1998) (function), JRip (Cohen, 1995) (rule-based) and Logistic model trees (Landwehr et al., 2005)(decision tree). Table 1 provides an overview of each algorithm applied on each of the 5 datasets with the corresponding f-measure. Numbers in bold indicate the dataset on which each algorithm best performed.The base feature set did better than the two baselines, increasing prediction quality by 2.7% and 4.1% respectively to 66.3%. Adding the UMLS to the base feature set (B+U) marginally increased the performance by 0.4%. The IAM feature set, when joined with the base (B+I), did better with 70.4%, an 1.4% increase. Finally, the dataset with the highest results is the Relief-F selected subset with a top scoring f-measure of 71.3%, which is an improvement of 0.9% over the 70.4% using the complete dataset (B+I+U), both attained with the voted perceptron algorithm.

5.3 Features relevance

The average absolute weight of each feature from the voted perceptron algorithm applied on the feature set B+I+U provided a ranking from the most discriminative feature to the least, for which the first 24 for each class are shown in Table 2. The first column, for the positive class (valuable), shows that the number of tokens still makes the top of the list with the five first features under different forms: number of content, any or stop tokens, percentage of similar (sim %) content tokens and number of sentence. The group (grp) and type (typ) UMLS features occupy almost half the list (10 out of 23) using similarity (sim) count, binary occurrence (bin) and number of occurrences (nbr). Five questions from the IAM questionnaire are also in the list, with three top ones being negative assessment from the physicians.

The second section shows for the negative class (non-valuable) that standard deviations (stddev) related to token length are the three most relevant features. 14 out of 24 features are from the UMLS resource with two-third (9 out of 13) using the similarity count. Three questions from the IAM questionnaire are also used. The rankings for the two output classes show that while length of comments is still a significant aspect of perceived value, features from the UMLS dataset are ranked high for their discriminative power. The similarity aspect is also used in half of the UMLs features

Table 2: Feature discriminative ranking from (B)ase, I(AM) and (U)MLS set

Positive	Negative
(B) Content Tok. [nbr]	(B) Content tok. [stddev]
(B) Tok. [nbr]	(B) Tok. [stddev]
(B) Stop Tok. [nbr]	(B) Stop tok. [stddev]
(B) Content tok. [sim %]	(B) Person pronoun [nbr]
(B) Sentence [nbr]	(B) Stop tok. [nom stddev]
(U) Typ embryo struct. [bin]	(U) Typ Occup [sim]
(U) Typ molecul. func. [sim]	(I) Reminded already knew
(B) Stop tok. [%]	(I) Learned something new
(B) Tok. [stddev nom *low*]	(B) Tok. [stddev nom *high*]
(I) I disagree with content	(U) Typ receptor [sim]
(B) LOE subtype	(U) Typ acid [sim]
(U) Grp Objects [nbr]	(U) Typ amino acid [bin]
(U) Typ bacterium [bin]	(U) Typ bacterium [sim]
(B) Tok. [stddev nom *avg*]	(U) Typ regul activ. [bin]
(U) Typ manuf object [nbr]	(B) Person pronoun [bin]
(I) There is a problem	(U) Typ receptor [sim]
(U) Grp Physiology [bin]	(U) Typ hazard subst. [sim]
(I) Not enough information	(U) Typ neoplas proc. [sim]
(U) Typ event [sim]	(U) Typ biomed occ. [bin]
(U) Grp Procedures [bin]	(U) Typ receptor [nbr]
(U) Typ bacterium [nbr]	(U) Grp Activ. Behav. [sim]
(I) Therapeutic approach	(I) Dissatisfied
(B) Summary structure [nbr]	(U) Typ eukaryote [bin]
(I) Info relevant for patient	(U) Grp Physiology [sim]

(11 out of all 23 UMLS features) which indicates the preponderant usefulness of this aspect over the other like binary occurrence and basic count.

An interesting observation on Table 2 is the type of IAM questions which prompt positive and negative value for the medical community. The top three IAM features used for the positive class (valuable) are from negative questions: ”‘I disagree with the content”’, ”‘There is a problem with this infoPoem”’, ”‘Not enough information”’. The negative class exhibit the same occurrence, as questions like ”‘I learned something new”’, ”‘Reminded of something I already knew”’ and ”‘I learned something new”’, which are supportive of the article, are used by the algorithm as highly discriminative features. This may suggest that comments shedding a negative view on article’s comments are considered more relevant than supportive ones. This would supports the brilliant-but-cruel hypothesis (Amabile, 1983).

5.4 Applicability

While the experiments results show good improvement over previous methods, enforcing the straight-out application of the trained classification model might not be advisable in the spirit of knowledge sharing in a continuing education program. As previously shown, the average inter-annotator agreement might be used to seek a more lenient classification method to select which comments are to be shared among the physician’s community and which are to be removed. One path to explore in this context is the opportunity that each algorithm can provide a confidence level which expresses the certainty of the algorithm regarding the chosen prediction class. It is usually based on the similarity rating between the features in the assessed entry and the ones in the trained model.

These results were generated using the voted perceptron algorithm. The dataset used for training was the 300 comments from the first step, combined with the 250 comments used for the inter-annotator agreement evaluation, using a majority vote to choose a relevant class. Finally, 450 randomly picked comments from the main dataset (3,470 comments) were added to these comments to provide a 1,000 comments dataset to the voted perceptron algorithm. The remaining 3,020 comments (3,470 minus the 450 retained for training) from the main dataset were used for evaluation purpose.

Table 3 shows the results for levels of confidence ranking from 75% to 95% for each individual class by providing the total number of comments classified as such, the number of errors (wrongfully classified comments) and the resulting prediction ratio. For example, the algorithm at a confidence level of 80% classifies 575 comments as being non-valuable. In these 575 comments, 65 were in fact classified by the annotator as belonging to the other class, resulting in a 88.7% success ratio, which is significantly higher than the best result from Table 1.

It can be observed in this table that while the non-valuable class provides a better success rate at the 95% confidence level, they all degrade to approximately 83% at 70% confidence rating. The non-valuable class shows better results but classifies a smaller amount of comments at the two higher confidence levels. It then drops to the same level for lower confidence than the valuable class, but progressively classifying more comments.

These new results could be used in two main scenarios for knowledge sharing among the medical practitioners. The first scenario is to use an arbitrarily chosen confidence level (for example, 80%) to filter out most of the non-valuable comments from the dataset. Physicians could then browse the remaining comments which would have a higher chance of being helpful. The sec-

Table 3: Precision performance per confidence level.

Confidence	Overall			Non-helpful			Helpful		
level	Precision %	Errors	Total	Precision %	Errors	Total	Precision %	Errors	Total
95	96.50	5	143	1	0	23	95.83	5	120
90	94,46	19	343	97,56	3	123	92,73	16	220
85	91,32	57	657	92,05	26	327	90,61	31	330
80	88,10	122	1,025	88,70	65	575	87,33	57	450
75	86,20	191	1,384	86,11	115	828	86,33	76	556

ond scenario is to use the confidence level as a ranking for all comments. Comments classified as valuable with a high confidence rating would be presented at the top of the list, followed by comments classified with a slightly lower confidence score and so on.

This second scenario provides more flexibility in an education context. The bottom of this list would be the top confidence scored comments for the non-valuable class, which could still be offered. This second scenario provides more flexibility in an education context. Even if most readers would only look at the top of the list, the curiosity driven readers would have access to the complete listing of comments which could be useful for topics relevant to their practice.

6 Discussion and Conclusion

Our applicative setting is one of information sharing in a context of continuing education for medical practitioners. This is certainly far from product review, but still the same problem exists that many comments are made by users, and these comments are not all useful to other users. Nevertheless, because of this applicative difference, performance comparison with other publications is not straightforward as we are not in a typical social media-based interactive setting, which means that typical data like star rating (used in Kim and Pantel (2006)) and relationship between reviews (like for Zhang et al. (2012)) is not available. Still, it can be observed in our performance evaluation that the basic unigram bag-of-word approach did not perform as well as our more complex features.

The UMLS features did not improve the overall performance when coupled with the base or the base+IAM questionnaire feature sets, probably because of the less relevant features which made data noisier. This is correlated with the increase of performance seen when the complete dataset was reduced to the 150 most discriminative features with the Relief-F feature selection algorithm. The IAM questionnaire, which physician are not obliged to answer completely, may suffer from the same problem of missing information on star rating for new products. Even if it is successfully used by classification algorithm, other features should be prioritized when possible. This could lead to more stable performances which are not dependent on the completeness of an external source of information.

As seen in the top negative features in Table 2, standard deviation was a useful measure for classification. The discretization of these features was also useful for both output classes. Although, while length can be a good predictor as shown in previous study (Kim and Pantel, 2006) and allows to discard useless short comments (”very good”, ”thanks”, etc.), it will undeniably wrongly classify comments like ”N?” (indicating the missing population size of a study) as useless. This is an extreme difficult case.

In conclusion, this research explored a textual feature extraction process with machine-learning classification to predict the helpfulness of comments in a context of continuing education for family physicians. Our research is well anchored in previous research on the topic, and we make further contributions by introducing similarity-based features to compare comments and infoPOEMs. We also introduce the use of an external domain specific resource to provide a measure of domain appropriateness for the comment, which is playing a role in its evaluation of helpfulness.

We showed that our method improved two previous baselines by 7.7% and 9.1% to a final 71.3% with the voted perceptron algorithm applied over a dataset of 150 features selected using the Relief-F algorithm. Since the categorization is far from perfect even if it gives good results (far above chance), we also suggested two confidence-based scenarios to make the categorization applicable in a real-world knowledge sharing context among medical practitioners.

References

TM Amabile. 1983. Brilliant but cruel: Perceptions of negative evaluators. *Journal of Experimental Social Psychology*, 19:146–156.

Hsinchun Chen and David Zimbra. 2010. AI and Opinion Mining. *IEEE Intelligent Systems*, 25(3):74–80, May.

William W. Cohen. 1995. Fast effective rule induction. In *Twelfth International Conference on Machine Learning*, pages 115–123. Morgan Kaufmann.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Danica Damljanovic, Thomas Heitz, Mark A Greenwood, Horacio Saggion, Johann Petrak, Yaoyong Li, and Wim Peters. 2011. *Text Processing with GATE (Version 6).*

Y. Freund and R. E. Schapire. 1998. Large margin classification using the perceptron algorithm. In *11th Annual Conference on Computational Learning Theory*, pages 209–217, New York, NY. ACM Press.

N. Friedman, D. Geiger, and M. Goldszmidt. 1997. Bayesian network classifiers. *Machine Learning*, 29(2-3):131–163.

Anindya Ghose and Panagiotis G. Ipeirotis. 2007. Designing novel review ranking systems: Predicting the Usefulness and Impact of Reviews. In *Proceedings of the ninth international conference on Electronic commerce - ICEC '07*, page 303, New York, New York, USA. ACM Press.

R Grad, P Pluye, and ME Beauchamp. 2006. Validation of a Method to Assess the Clinical Impact of Electronic Knowledge Resources. *E-Service journal.*

RM Grad, P Pluye, and J Mercer. 2008. Impact of research-based synopses delivered as daily e-mail: a prospective observational study. *Journal of the American Medical Informatics Association.*

Roland Grad, Pierre Pluye, Vera Granikov, Janique Johnson-Lafleur, Michael Shulha, Soumya Bindiganavile Sridhar, Jonathan L. Moscovici, Gillian Bartlett, Alain C. Vandal, Bernard Marlow, and Lorie Kloda. 2011. Physicians' assessment of the value of clinical information: Operationalization of a theoretical model. *Journal of the American Society for Information Science and Technology*, 62(10):1884–1891.

M Hall, E Frank, G Holmes, and B Pfahringer. 2009. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1).

Jina Huh, Meliha Yetisgen-Yildiz, and Wanda Pratt. 2013. Text classification for assisting moderators in online health communities. *Journal of Biomedical Informatics*, (0):–.

SM Kim and Patrick Pantel. 2006. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, number July, pages 423–430.

JR Landis and GG Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics.*

Niels Landwehr, Mark Hall, and Eibe Frank. 2005. Logistic model trees. *Machine Learning*, 59(1-2):161–205.

Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz, and Timm Euler. 2006. YALE: Rapid Prototyping for Complex Data Mining Tasks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06*, page 935, New York, New York, USA. ACM Press.

Arjun Mukherjee and Bing Liu. 2012. Modeling Review Comments. In *Proceedings of 50th Anunal Meeting of the Association for Computational Linguistics.*

Thomas L. Ngo-Ye and Atish P. Sinha. 2012. Analyzing Online Review Helpfulness Using a Regressional ReliefF-Enhanced Text Mining Method. *ACM Transactions on Management Information Systems*, 3(2).

John C. Platt. 1998. Sequential minimal optimization: A fast algorithm for training support vector machines.

Pierre Pluye, Roland M. Grad, Vera Granikov, Justin Jagosh, and Kit Leung. 2010a. Evaluation of email alerts in practice: Part 1. Review of the literature on clinical emailing channels. *Journal of Evaluation in Clinical Practice*, 16(6):1227–1235, December.

Pierre Pluye, Roland M. Grad, Janique Johnson-Lafleur, Tara Bambrick, Bernard Burnand, Jay Mercer, Bernard Marlow, and Craig Campbell. 2010b. Evaluation of email alerts in practice: Part 2 - Validation of the information assessment method. *Journal of Evaluation in Clinical Practice*, 16(6):1236–1243.

H Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

PD Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, (July):417–424.

Wenting Xiong and Diane Litman. 2011. Automatically predicting peer-review helpfulness. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, number 2009, pages 502–507.

Richong Zhang and Thomas Tran. 2008. An Entropy-Based Model for Discovering the Usefulness of Online Product Reviews. *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 759–762, December.

Kunpeng Zhang, Yu Cheng, Wei-keng Liao, and Alok Choudhary. 2012. Mining millions of reviews: a technique to rank products based on importance of reviews. In *Proceedings of the 13th International Conference on Electronic Commerce - ICEC '11*, pages 1–8, New York, New York, USA. ACM Press.