Sentiment Independent Topic Detection in Rated Hospital Reviews

Christian Wartena
Hochschule Hannover
Christian.Wartena@hs-hannover.de

Uwe Sander
Hochschule Hannover
Uwe.Sander@hs-hannover.de

Christiane Patzelt
Hochschule Hannover
Christiane.Patzelt@hs-hannover.de

Abstract

We present a simple method to find topics in user reviews that accompany ratings for products or services. Standard topic analysis will perform sub-optimal on such data since the word distributions in the documents are not only determined by the topics but by the sentiment as well. We reduce the influence of the sentiment on the topic selection by adding two explicit topics, representing positive and negative sentiment. We evaluate the proposed method on a set of over 15,000 hospital reviews. We show that the proposed method, Latent Semantic Analysis with explicit word features, finds topics with a much smaller bias for sentiments than other similar methods.

1 Introduction

There are many websites that collect user opinions and ratings on products or reviews. In this paper we study a collection of reviews and ratings of orthopedic treatments in hospitals. On the leading German social media website for hospital rating www.klinikbewertungen.de users may rate and comment about 3000 hospitals. On this website, in principle it is possible to see what topics are criticized and which ones are valued. To do so, we need to do a topic analysis on the comments.

Since many texts in our corpus have a strong polarity, a standard topic analysis, using Probabilistic Latent Semantic Analysis (PLSA) or Latent Dirichlet Allocation (LDA) also tries to account for the words associated with positive and negative sentiment. Most likely topics and sentiments will be mixed up. E.g. the topic pain is usually associated with negative feelings. Thus negative opinion and pain get mixed up in one topic. Consequently, the topic *Pain* might be found for a document that contains negative words but is not about pain and, vice versa, a document talking about a positive experience on pain treatment will not be associated with the topic *Pain*. Thus we have to model the sentiment and the topic independently.

A straightforward way to make the topic analysis sentiment independent would be to treat comments that come with positive and those that come with negative ratings separately. However, we would end up with incomparable topics for positive and negative comments. Joint topic-sentiment models are designed to find topics and polarity of each document, while we already have the polarity of each document. Moreover, these models are designed to optimize sentiment analysis and not to make the topics less biased towards some sentiment.

The solution we present in Section 2 is basically a simplified formulation of the method proposed by Mei et al. (2007). We use Latent Semantic Analysis (LSA) and add fixed topics for positive and negative sentiment to the set of topics that have to be learned. Thus much of the positive and negative words are explained by these dimensions and less of these words are explained by the other topics.

2 Method

In order to keep the influence of positive or negative opinion out of the topic modeling, we add two fixed topics representing these sentiments to the LSA model. These topics are initialized with values calculated before and not updated in the learning phase. As values for these fixed dimensions we either take the ratings for each document or we compute the polarity of each word.

LSA (Landauer and Dumais, 1997) is a simple but effective method for topic analysis: a term-document matrix is decomposed into two smaller matrices. The rows of the first matrix can be interpreted as the topic distributions of the documents while the second matrix gives the word distribution for these topics. The decomposition is usually realized by Singular Value Decomposition. In the following we will use Non-negative Matrix Factorization (NMF) (Pentti and Unto, 1994) for the decomposition, which makes the weights easier to interpret and can be seen as a variant of PLSA (Hofmann, 2001; Gaussier and Goutte, 2005). We start with the term-document matrix TD of size $m \times n$, with m the number of documents and n the number of terms. Each element $TD_{i,j}$ is the weight of word j for document i. Now we assume that there are k (latent) topics (with $0 < k \ll n$) such that TD can be decomposed into a document-topic matrix U of size $m \times k$ and a word-topic matrix V of size $n \times k$ Since we do not know the topics, we choose some k, initialize U and V randomly and use the stochastic gradient descent algorithm to minimize $||TD - U \cdot V^T||_{Fro^2}$. Furthermore, we require that the row vectors of V^T have magnitude 1.

2.1 LSA with explicit features

As fixed dimensions for positive and negative sentiment we can directly use the given ratings. We initialize the first two columns of U with these values and we will never update these values in the optimization process. Formally, we set

$$
\begin{aligned}
U_{i,0} &= max(0, r_i - 1) \text{ and} &(1)\\
U_{i,1} &= max(0, 2 - r_i) &(2)
\end{aligned}
$$

for each $0 < i \leq m$ where $r_i \in \{0, 1, 2, 3\}$ is the rating associated with document i. We use seperate columns for positive and negative sentiment, since negative sentiment is not just the absence of positive sentiment and neutral words or documents are not words or documents somewhere inbetween positive and negative sentiment, but they are lacking this dimension. We call this method LSA with explicit document features (LSA-ExplDF).

Alternatively, we first determine the positive and negative polarity of each word and initialize the first two columns of V. For this purpose we compute the Information Gain (IG) of every word for the probability function that a document has a positive rating: Let P be a discrete random variable with values 0 and 1 indicating the polarity of the document. Now $H(P)$ is the entropy of P and $H(P \mid w)$ is the relative entropy of P given that it is known whether the word w is in the document. The IG of w is now defined as $I(w) = H(P) - H(P \mid w)$. Finally we set

$$
V_{i,0} = \begin{cases} I(w_i) & \text{if } df(C_{pos}, w) > df(C_{neg}, w) \\ 0 & \text{otherwise} \end{cases} \tag{3}
$$

$$
V_{i,1} = \begin{cases} I(w_i) & \text{if } df(C_{neg}, w) > df(C_{pos}, w) \\ 0 & \text{otherwise} \end{cases} \tag{4}
$$

where $df(C_{neg}, w)$ is the relative document frequency of w in set of all negatively rated documents and $df(C_{pos}, w)$ the relative document frequency of w in set of all positively rated documents. We call this variant LSA with explicit word features (LSA-ExplWF). We implemented the algorithms in Python using the Stochastic Gradient Descent method for NMF from the Scikit-learn package (Pedregosa et al., 2011).

In the following we use LSA-ExplWF and LSA-ExplDF to force the factorization to have dimensions related to positive and negative sentiment. However, we could use the same method for any aspect from the corpus that we want to be represented explicitly, and that should not influence the other topics, like e.g. genre or style.

3 Data

We have tested various topic detection algorithms on reviews from the German platform for hospital reviews, `www.klinikbewertungen.de` (Drevs and Hinz, 2014). Each review consists of five satisfaction scales (overall satisfaction, quality of consultation, medical treatment, administration and procedures, equipment and designing of structures, on a 4-point scale), optional fields for comments about the hospital stay, the pros and cons and the disease pattern in their own words. For our study, in March 2016, we retrieved all reviews for the orthopedics departments. Data collection includes 15,840 reviews of 14,856 patients (93.8%), 852 relatives (4.7%), 30 clinicians/doctors/hospital staff (0.1%) and 102 other affected persons (0.6%) from 1072 hospitals and rehabilitation facilities. 12,098 (76.4%) reviews have positive and 3,742 (23.6%) negative overall rating.

Since in most cases only one text field is filled, we only used the overall rating and concatenated all text fields to one single text. The total number of words in the corpus of these texts is 2,489,356. To construct the term document matrix we lemmatize all words using the Tree Tagger (Schmid, 1994), compute the document frequency and select all nouns and verbs that occur in more than 5 documents and in less than half of all documents. This results in a list of 7,596 words that we use to represent each document. We did not include adjectives and adverbs. These words may also bear topical information but these words are used very frequently to express the sentiment. By excluding these words, we remove already a lot of sentiment from the documents and make the topic detection more neutral with respect to sentiment.

The values of the term-document matrix are the tf.idf values for each term and document. For a corpus C, each term t and document $d \in C$ we define $\mathit{tf.idf}(C, t, d) = \mathit{tf}(t, d) \cdot \mathit{idf}(t)$ with

$$\mathit{tf}(t,d) = 1 + \log\left(1 + \frac{n(t,d)}{\max_{t' \in V} n(t',d)}\right) \tag{5}$$

$$\mathit{idf}(t,C) = \log\left(1 + \frac{|C|}{\mathit{df}(C,t)}\right) \tag{6}$$

where V is the vocabulary of C, $n(t,d)$ is the number of occurrences of t in d and $\mathit{df}(C,t)$ is the number of documents in which t occurs. We minimize the effect of the term frequency since the documents that we consider are concatenations of different fields and therefore some words occur more frequently only because it related to an aspect that was asked for in more than one field.

4 Evaluation

We compare three variants: LSA (with NMF), LSA-ExplDF and LSA-ExplWF. In all cases we set the number of topics k to 20 plus the number of fixed topics.

Since the goal is to make the topics independent of the sentiment, we will use exactly this as an evaluation criterion. For each document we determine the two most prominent topics, assuming that there are at least two topics in each text. The results, however, do not depend on the number of topics chosen. Subsequently, we count the number of times each topic is assigned to a negative and to a positive document. If the topics would be completely independent, the ratio of positive and negative documents would be the same for each topic. Thus we take the variance of the fraction of negative documents for each topic as criterion for success: the lower the variance the more independent the topics are from the sentiment. Of course the topics are not independent of the sentiment. Nevertheless a smaller variance indicates that topics ans sentiments are better seperated. Since the results are not deterministic we use averages of 10 runs.

Table 1 gives the average fraction of negative documents and the variance for each method. A lower variance shows that the fraction is more similar for each topic, indicating that the topics are more independent of the sentiment of the texts. We clearly see that LSA-ExplWF give the best result and impressively reduces the variance between the percentage of negative document per topic.

Table 1: Fraction of negative documents per topic.

Method	Average	Variance
LSA	0.25	0.49
LSA-ExplDF	0.26	0.46
LSA-ExplWF	0.25	0.17

Table 2: Most prominent words for all topics found by LSA-ExplWF. Fixed topics are excluded. See Table 3, second column for typical words for the first two topics.

1	sagen, .., tun, gehen, wissen	11	nehmen, zeit, frage, beantworten, erklären
2	reha, therapeut, anwendung, essen, zimmer	12	kind, kur, mutter, kinderbetreuung, tochter
3	operieren, op, operation, hüfte, dr.	13	frau, herr, dr., dank, dr
4	station, schwester, krankenhaus, op, pflegepersonal	14	betreuung, verpflegung, versorgung, unterbringung, behandlung
5	umgebung, schwimmbad, wochenende, nutzen, ort	15	therapie, therapeut, therapieplan, abstimmen, servicepersonal
6	patient, mitarbeiter, freundlichkeit, aufenthalt, kompetenz	16	lws, hws, bandscheibenvorfall, bws, schmerz
7	termin, schmerz, wartezeit, untersuchung, mrt	17	knie, kniegelenk, arthrose, op, tep
8	zimmer, fernseher, internet, telefon, tv	18	tep, hüft, hüfte, ahb, gehhilfen
9	frühstück, abendessen, auswahl, mittagessen, salat	19	wunsch, eingehen, erfüllen, bedürfnis, berücksichtigen
10	massage, vortrag, übung, anwendung, gruppe	20	nicht, und, war, ich, mit

To get an impression of the topics found, Table 2 gives the five most prominent words for each topic found by one run of LSA-ExplWF.

Though the results differ slightly across two runs, most topics are found in each run and many topics found by one method also are found by another method. E.g. both methods find a topic that can be represented by the words *Therapie* (therapy), *Therapeut* (therapist), etc. (topic 15 in Table 2). In the case of LSA this topic was assigned to 888 positive and 569 negative documents. In LSA this topic thus has a strongly negative connotation. Using LSA-ExplWF the topic was assign to 978 positive and 373 negative documents. The comment *"Ich habe mich hier ausgesprochen wohl gefühlt, als ich eine künstliche Hüfte (TEP) erhalten hatte und nach dem Krankenhausaufenthalt drei Wochen in dieser Reha Klinik verbrachte. …Die Therapie wurde ganz individuell auf meine Bedürfnisse abgestimmt. … "* (I felt very well here when I got an artificial hip (TEP) and spent three weeks in the rehabilitation clinic after the hospital stay. …The therapy was individually tailored to my needs. …) got topics 15 and 18 from LSA-ExplWF, while LSA assigned topics 2 and 18, probably because topic 15 has a negative bias in LSA and did not fit for this positive comment. In another example a patient is massively complaining that the doctors did not take time for him, that there was only a standard treatment and he could not shower every day. Here LSA assigns topics 15 and 1, while LSA-ExplWF assigns topics 11 (taking time, answering questions) and 14 (nursing care). LSA probably assigns topic 15 mainly because the text is extremely negative, while LSA-ExplWF precisely identifies the topics that frustrated the patient.

Table 3 gives the most prominent words for the first two topics. Interestingly, the word *empfehlen)* (recommend) is found both for positive and negative sentiment: this word is used in stronly polarized contexts, both with positive and with negative sentiment, but it is not used frequently in neutral reviews.

5 Related Work

Much work on topic detection in combination with sentiment analysis was done on product reviews. The semi-supervised model of McAuliffe and Blei (2008) optimizes the topics for rating prediction. Besides the rated products there are aspects of these products that are discussed positively and negatively. Titov

Table 3: Positive and negative words in LSA-ExplDF and LSA-ExplWF.

	LSA-ExplDF	LSA-ExplWF
+	personal *(staff)*	team
	empfehlen *(recommend)*	dank *(thanks)*
	essen *(meal)*	fühlen *(feel)*
	top	bedanken *(to thank)*
	super	aufheben *(to save)*
-	arzt *(doctor)*	katastrophe
	katastrophe	aussage *(statement)*
	patient	ignorieren *(to ignore)*
	empfehlen *(recommend)*	nachfrage *(demand)*
	geld *(money)*	geld *(money)*

and McDonald (2008) use the same type of data we have. They consider the problem that ratings are given on several aspects but only one textual comment is given. This is also the case for our data. However, they distinguish between global topics and local topics that correspond to ratable aspects of the global topics. They propose an extension of Latent Dirichlet Allocation (LDA) to handle this mixture of global and local topics. In our data, that are much more specific, we did not find such a division between global and local topics and the global topics correspond very well to ratable aspects. Zhao et al. (2010) propose an extension of this model that is able to use various features of words and can distinguish aspect from opinion words.

Much work was done on developing joint topic-sentiment models, usually to improve sentiment detection. Lin and He (2009) propose a method based on LDA that explicitly deals with the interaction of topics and sentiments in text. However, their goal is exactly opposite to ours: they use the fact that the topic distribution is different for positive and negative documents and in fact use the polarity of topics to enhance the sentiment detection, which is the main goal of their efforts. Thus the algorithm is encouraged to find topics that have a high sentiment bias. The joint topic sentiment model of Eguchi and Lavrenko (2006) goes into the same directions: they optimize sentiment detection using the fact that the polarity of words depends on the topic. Also the paper of Maas et al. (2011) follows this general direction. Paul and Dredze (2012) propose a multidimensional model with word distributions for each topic-sentiment combination. This model was used to analyze patient reviews by Wallace et al. (2014).

The work of Mei et al. (2007) is most similar to our approach. In fact our method can be interpreted as a simplification of their method. A difference is that Mei et al. use a background word distribution that is topic and sentiment independent to account for general English words. We also tried this, but such a component did not have any effect. This can be explained by the fact that we removed stop words and used tf.idf weights instead of raw counts. One of their goals also is to avoid a contamination of topics with sentiments. However, they did not evaluate this aspect. Thus the contribution of this paper is not just a simpler formulation of the basic idea of Mei et al. (2007), but also shows that the topics found indeed are less contaminated by sentiment words and less biased towards one sentiment.

6 Discussion

The proposed method gives a simple but effective way to find topics in strongly polarized texts if the polarity of the texts is known, as usually is the case in comments given in rating portals. We have shown on a realistic data set, that the topics found become more independent from the sentiment. We could also show the effect of our method on a few example texts.

Patient comments often have different opinions on different topics. For future work we will try to find out for each comment the topics it is discussing positively and negatively.

References

Drevs, F. and V. Hinz (2014). Who chooses, who uses, who rates: The impact of agency on electronic word-of-mouth about hospitals stays. *Health care management review 39*(3), 223–233.

Eguchi, K. and V. Lavrenko (2006). Sentiment retrieval using generative models. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 345–354. Association for Computational Linguistics.

Gaussier, E. and C. Goutte (2005). Relation between plsa and nmf and implications. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, New York, NY, USA, pp. 601–602. ACM.

Hofmann, T. (2001, Jan). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning 42*(1), 177–196.

Landauer, T. K. and S. T. Dumais (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review 104*(2), 211.

Lin, C. and Y. He (2009). Joint sentiment/topic model for sentiment analysis. In *CIKM*.

Maas, A. L., R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, Stroudsburg, PA, USA, pp. 142–150. Association for Computational Linguistics.

McAuliffe, J. D. and D. M. Blei (2008). Supervised topic models. In *Advances in neural information processing systems*, pp. 121–128.

Mei, Q., X. Ling, M. Wondra, H. Su, and C. Zhai (2007). Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, New York, NY, USA, pp. 171–180. ACM.

Paul, M. and M. Dredze (2012). Factorial lda: Sparse multi-dimensional text models. In *Advances in Neural Information Processing Systems*, pp. 2582–2590.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research 12*(Oct), 2825–2830.

Pentti, P. and T. Unto (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics 5*(2), 111–126.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, Manchester, UK, pp. 44–49.

Titov, I. and R. McDonald (2008). A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL-08: HLT*, pp. 308–316. Association for Computational Linguistics.

Wallace, B. C., M. J. Paul, U. Sarkar, T. A. Trikalinos, and M. Dredze (2014). A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. *Journal of the American Medical Informatics Association: JAMIA 21*(6), 1098–1103.

Zhao, W. X., J. Jiang, H. Yan, and X. Li (2010). Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, Stroudsburg, PA, USA, pp. 56–65. Association for Computational Linguistics.