Topic Detection in Noisy Data Sources*

Kerstin Denecke
L3S Research Center
Hannover, Germany
denecke@L3S.de

Marko Brosowski
L3S Research Center
Hannover, Germany
brosowski@L3S.de

Abstract

Automatic topic detection becomes more important due to the increase of information electronically available and the necessity to process and filter it. In particular, when language is noisy like in weblog postings, it is challenging to determine topics correctly. Nevertheless, it is still unclear, to what extent existing topic detection algorithms are able to deal with this noisy material. In this paper, Latent Dirichlet Allocation (LDA) is exploited to determine topics in weblog sentences. We perform an extensive evaluation of this algorithm on real world data of different domains. The results show that LDA can successfully determine topics even for short and noisy sentences.

1 Introduction

The need to have methods on hand that enable automatic analysis and interpretation of Web data becomes more and more important. This is due to the increase of unstructured data available. For lots of applications that rely upon Web data, detecting topics of documents on different levels of granularity is necessary. Consider the following example: Given a medical weblog post dealing with a specific medical treatment. The author highlights different aspects of this treatment and presents arguments in favor and against it. In order to collect and summarize these arguments, it is of substantial interest to identify sentences that are dealing with the same aspect and summarize the expressed opinion. It is insufficient to only know the general topic of the complete post.

In order to enable such applications, topics need to be detected for different document granularity (document-level, paragraph-level, sentence-level). In this paper, the focus is on detecting topics on sentence-level. Among others the previously mentioned application requires normally this level of granularity.

*This work has been done within the LivingKnowledge project, partly funded by the European Commission under 231126.

A frequently cited algorithm for topic detection is topic modeling by Latent Dirichlet Allocation (LDA). So far, it has been mainly applied for topic detection at document-level [2]. The quality of LDA for detecting topics remained unconsidered until now. We will go a step further, by applying LDA to sentence-level topic detection for weblog posts and studying the LDA quality on sentences. This problem is even more difficult, since for single sentences context information is missing. Further, it is still unknown, how the quality of LDA results is influenced by the pre-selected number of topics to be distinguished. In this paper, a comprehensive evaluation is performed that also studies the label quality of LDA depending on the number of clusters to be distinguished.

The main contributions of this work can be summarized as follows: (1) Adaptation of LDA to sentence-level topic detection, (2) an extensive evaluation of LDA for topic detection in noisy sentences of three different Weblog datasets, and (3) an extensive assessment of correlations of the number of topics and the quality of topic detection.

2 Related Work

Approaches to topic representation and detection can be clustered in approaches for keyword or keyphrase extraction, lexicon lookup, and topic modelling. Keywords and -phrases are characteristic words and phrases used in a document. Supervised and unsupervised algorithms have been exploited for keyword extraction [11]. But, these approaches are limited to the extraction of known keywords. The TopCat system [4] exploits natural language processing techniques to identify key entities in texts and then forms clusters with a hypergraph partitioning scheme. KEA [5] identifies keyphrases using TFxIDF and a Naive Bayes classifier. Unsupervised algorithms include string frequency or feature weight calculation [7] and semantic network structure analysis [6].

For detecting topics in product or movie reviews, substituting topic identification with a lexicon look-up to determine product names, person names and the like as topics

within the opinion mining task has been proven successful [9], [10].

Topic models as introduced by Blei et al. consider documents as mixture of topics [2]. Each topic is represented by a set of keywords together with a probability indicating the word's contribution to the topic. Within clustering approaches such as Latent Dirichlet Allocation [2] automatic labeling of clusters is difficult.LDA has been used in different application scenarios besides topic detection, e.g., for identifying Spam Web sites [1], detecting frauds in telecommunication [12] or assessing sentiments expressed towards topics [8].

In this work, noisy sentences are taken into account as subject of interest in topic detection. This makes the problem even more complicated since within a sentence, a topic might not always be mentioned explicitly. For this reason, keyword extraction and lexicon lookup approaches to determine topics are unsuited in our context. Further, Weblogs postings can deal with very different topics which makes creation of an appropriate lexicon difficult. For these reasons, LDA is chosen as approach to topic detection since it is suited best for the problem and data we are addressing. Blei et al. [2] evaluated the LDA algorithm on document-level aiming at estimating the density of topic models. In particular, the perplexity of a held-out test set is calculated to evaluate the topic models. Instead, we will focus on an extensive quality assessment of LDA for the task of sentence-level topic detection. Results of LDA quality for topic detection in documents in general, and in sentences of noisy Web data in particular, are still unavailable.

3 Identification of Topics

In this paper, we are considering the theme of a sentence as its 'topic' which can be described by a set of words that do not have to be explicitly mentioned in a sentence. We determine these topic describing terms (referred to as 'topic terms') using LDA. The original Blei implementation of LDA is extended by a sentence detection algorithm to apply it to sentences. The complete process comprises five steps described in more detail in the following paragraphs. Example topics with sentences are shown in Table 1.

Document splitting First, each post is split into sentences using the sentence splitting library provided by Lingpipe[1]. Lingpipe is a framework for linguistic analysis of human language that offers java libraries for text classification and linguistic processing. LingPipe also corrects bad formatted endings of sentences including sentences without proper space, dots, and question marks.

[1] http://alias-i.com/lingpipe/

Sentence and Word Normalization In a second step, the sentences are normalized: Only nouns and proper nouns are stemmed and kept for further processing. For detecting word classes and to perform stemming, the Stanford NLP Toolset[2] is used. We restrict the words to be considered by LDA to nouns to reduce computing time and to restrict topic terms to those that are content-bearing. Our previous experiments showed that for long sentences and huge datasets, the calculation of topic models with LDA is very time consuming. Further, verbs, adjectives and adverbs are often unsuited to describe the topic of a sentence.

Topic Detection In a third step, topics along with their probabilities are identified for each sentence using the LDA-algorithm and based on the vector representation of (normalized) sentences. Each sentence is considered to consist of a topic mixture and each word's creation is attributable to one of the sentence's topics. Therefore, a topic is described by a set of words derived from the documents where to each word a probability is assigned that indicates the relevance of this word for the topic. In this way, all topics are described by the same words, but with varying probability values for each word. The output of LDA is finally (1) the probability of each word for a topic and (2) the probability of each topic for a sentence. Through previous experiments we learned that it is necessary to pre-filter sentences regarding their topical focus. In order to exclude sentences without topical focus, our LDA modification considers only sentences with at least four words (excluding stop words). A term is in turn considered relevant for clustering when it occurs in at least 15 sentences.

To run LDA, the number of clusters to be formed needs to be fixed. Since it is still unknown what the best number of clusters to be chosen is, we perform an extensive evaluation to identify correlations between the number of topic clusters and the data set size. LDA also requires to fix two other parameters: The α parameter determines how dominant a topic is going to be in a document. The hyperparameter β can be interpreted as the prior observation count on the number of times words are sampled from a topic before any word from the corpus is observed. Steyvers and others have found $\alpha = \frac{50}{t}$ and $\beta = 0.01$ where t is the number of topics to work well with many different text collections [3]. We chose these values in our experiments.

Topic Selection In a last step, the probabilities determined by LDA are used to select the topic and topic terms for a sentence. The probability per topic and sentence calculated by LDA indicates to what degree the sentence belongs to the topic. If for a sentence all topics have the same probability, they are equally distributed with a probability of $\frac{1}{k}$

[2] http://nlp.stanford.edu/software/tagger.shtml

correct	Text
WebMD (k=20), Topic "eye cancer surgery ear vision"	
yes	LASIK flap trauma can cause the flap to completely come off the eye ... bad news !
yes	Visual conflicts include crooked eyes (strabismus) , blockage of the normal visual pathway (as in a dense infantile cataract) , or a marked difference in the refractive power of the two eyes (as an example : one eye nearsighted and the other farsighted)
no	How do you handle their concern and their anxiety about going swimming ?
Reviews (k=50), Topic "cd software installation time"	
yes	I work on word, explorer, netscape, acrobat reader, photoshop at same time (meaning using a lot of memory), but still dont see a problem with computer (getting slow or crashing).
yes	The setup wasn't quite perfect the first time – I needed to go through a web-based interface to set up the router (to give it info on my DSL provider account), though the documentation for this was staightforward.
no	the pictures come out hazy if your hands shake even for a moment during the entire process of taking a picture .
Slashdot (k=200), Topic "cable market competition service monopoly "	
yes	But perhaps simply the threat of competition will get things moving again .
yes	The legislation that will allow for national cable TV franchises will not be passed into law or at least a year (if at all).
no	what you claim can't really be seen from satelite pictures that were shown as supposed proof .

Table 1. Example sentences with detected topics.

(with k = number of topics). In this case, the probability does not allow to draw any conclusions on the most probable topic of a sentence. For this reason, in our approach, topics with a probability larger than $\frac{1}{k}$ are considered as topics of a sentence. Topics with a smaller probability are excluded since their support for describing the content of a sentence is too low. In this way, up to three topics are selected for each sentence. One reason for choosing LDA are these probability values that allow us to filter out irrelevant topics and also sentences without topical focus.

Topic Term Selection LDA also provides the probabilities of each word for a topic. Since we want to assess the quality of LDA topic detection, we need terms that describe each topic. Our preliminary studies showed that the five most probable words are suited best for describing the topic of a sentence. Therefore, we are considering the five words with the highest probability for a topic as 'topic terms'. They are used to characterize a topic. We will report results of a study that assesses the feasibility of these five terms for describing a topic later in this paper.

As outlined in the description before, the LDA algorithm is not looking for matching keywords, but it is creating a model that describes to what extent the single words of a dataset are relevant for single topics (or clusters) of sentences. The benefit of these topic models is that the correct topic can be assigned even if no matching keyword occurs in the sentence, just by relying upon a larger set of words, or the context, respectively. Therefore, it can happen that none of the top 5 highest ranked topic describing words are included in the document under consideration.

4 Experimental Evaluation

In this section, we describe the evaluation methodology, the real-world datasets used in our study, and present the evaluation results.

4.1 Evaluation Methodology and Objectives

The objectives of this evaluation are to assess the: (1) correctness of LDA in assigning topics at sentence level, (2) correlation between the quality of LDA and the number of pre-selected topics, and (3) dependency of quality related to length of sentences. We will also assess the annotator agreement in deciding for the correctness of an assigned topic. Further, characteristics of the data sets that can be derived from the LDA results are reported.

Two people were involved in the assessment of topic quality. One of them is author of this paper. They were confronted with the five most probable topic terms of the topics and had to select the group of terms that describes best the topic of the sentence under consideration. If no topic was suited they were asked to indicate this. The evaluators were asked to consider a topic correct even if it describes the sentence topic in an abstract way. At least one suggested topic word had to describe the content of the sentence and the topic has to be selected as correct where most of the topic terms are feasible. Since we intend to study the correlation between number of topics, quality of topic detection and data set size, we performed experiments where LDA has to distinguish 20, 50, 100, 150, or 200 topics. The annotators were confronted with each sentence for five times, i.e. for each sentences and five different sets of assigned topics, they had to select the one that described the sentence content best.

The manual evaluation is performed on 3000 selected sentences (see next section) whereas the study of the characteristics of the datasets considers the complete data collection. For different numbers of topics, the accuracy value is determined per dataset to measure the quality of topic detection. A sentence is considered 'correct' if the annotator selected one assigned topic; otherwise the sentence is considered 'false'. Further, the annotator agreement in this task is determined.

In a second evaluation, the consistency of topic terms is assessed, i.e., the annotators judge to what extent the five topic terms that have been selected for describing a topic, deal with the same subject (e.g., the topic terms *hair, computer, drug, company, year* are not describing the same topic while *disease, blood, heart, risk, change* clearly describe one topic). The annotators were confronted with the five topic terms per topic and were asked to mark the words that do not fit together. Based on these annotations, the consistency of a topic is measured by calculating the number of topic words that are not describing the same topic. A topic is consistent if at least 4 of the 5 topic terms fit together. It is slightly consistent, when two or three topic words belong together. Otherwise, the topic is considered inconsistent.

For example, the topic with topic terms *child, parent, kid, school, boy* is considered 'consistent' since all terms terms describe the same topic. In contrast, the topic described by the terms *condition, behaviour, approach, computer, expert* is considered 'slightly consistent' because two words fit together at most.

4.2 Corpus Description

The modified LDA is applied to three different data sets. The first data set (referred to as WebMD) consists of 28 health-related blogs with 2,405 posts covering 4 years (January 2005 to January 2009) collected from the WebMD webpage (http://blogs.webmd.com/). The second data set (referred to as Slashdot) comprises comments from the Slashdot website. Slashdot is a popular website for people interested in reading and discussing about technology and its ramifications. It includes posts, as well as comments on these posts. In this study, we used a dataset provided for the CAW2 workshop (http://caw2.barcelonamedia.org/) that contains about 140,000 comments under 496 articles, covering the time period from August 2005 to September 2006. We consider only the comments, not the posts themselves.

The third dataset (referred to as Reviews) is given by reviews on 14 products collected from Amazon.com. The reviews deal with the following products: cellular phones, MP3 player, DVD player, digital cameras, diaper, router, and security software. This data set is provided by Minqing Hu and Bing Liu (http://www.cs.uic.edu/l̃iub).

We are considering sentences of our data sets as 'noisy' since they contain writing errors, common speech or even slang, incomplete sentences and unknown abbreviations (e.g., *Hehehe..... suffice it to say, erm, DHS is quite a similar experience.... but I'm not one to gossip, so you didn't hear that from me*). These linguistic characteristics make the processing more complicated.

Since we are analysing the quality of topic detection at sentence-level, sentences that are understandable by humans even without seeing the context are required. For this reason, we created a reasonable sample of the three corpora by confronting two reviewers with randomly selected sentences of the data sets. The annotators had to decide whether they understand the sentence. In this way, for each data set 1000 understandable sentences were selected. A sentence was added to the sample if both annotators agreed that they understand the sentence. This sample is used to study the quality of LDA in our experiments.

4.3 Quality of Topic Detection

In this section, we report the quality assessment results of topics assigned to sentences using the modified LDA.

	k=20	k=50	k=100	k=150	k=200
WebMD	86.6% (650)	82.9% (677)	74.1% (731)	70.7% (690)	70.6% (671)
Reviews	86.7% (732)	85.6% (793)	88% (699)	87.3% (645)	89.4% (565)
Slashdot	62% (565)	54.2% (606)	60.8% (671)	62.7% (673)	64.7% (665)

Table 2. Accuracy for the different datasets (with k as number of topics) and total number of sentences where both annotators agreed.

Annotator Agreement Depending on the number of topics LDA had to distinguish, the annotators agreed in between 65% and 73.1% of the sentences for WebMD. The agreement was slightly better for sentences of the customer review data set (agreement for 79.3% of the sentences). For topics assigned to the Slashdot dataset, the test persons disagreed to the largest extent. The largest agreement throughout all three datasets was achieved when LDA distinguished 50 or 100 topics. One reason for this might be that the topics become more overlapping when sentences have to be assigned to 150 or 200 clusters. The annotators have then difficulties to decide for one of the assigned topcis. Additional reasons for annotator disagreement are described in the discussion section. Anyway, the annotator agreement results show that topic detection is a difficult task, even for humans and that opinions regarding the correctness of an assigned topic can differ. In the following sections the reported accuracy results have been calculated by only considering the annotations were the annotators agreed.

Accuracy of Topic Detection Depending on the number of topics and the dataset, accuracy values between 54.2% and 89.4% are achieved (see Table 2). The topic detection achieved the worst results for the Slashdot dataset with up to 54.2% accuracy. In contrast, the evaluation of the customer review dataset resulted in significantly better accuracies between 85.6% and 89.4%. Only sentences were both annotators agreed were considered which leads to different numbers of sentences considered for the several values of k.

Given the values of k where the annotators agreed to the largest extent, the accuracies of topic detection are 74.1% for WebMD (k=100), 85.6% for the reviews (k=50) and 62.7% for the Slashdot dataset (k=150). It can be seen, that topic detection in customer reviews performs significantly better than in the other corpora. One reason might be that each review clearly deals with exactly one product: Product features such as *battery, display* which can be seen as topics are mentioned explicitly. In contrast, topics in sentences of WebMD or Slashdot postings are sometimes only subtly described which makes topic detection more difficult.

We further studied how many topics have been assigned to a sentence and which of them were labeled correct in our evaluation. For the WebMD and Slashdot dataset, between

45-60% of the sentences had only one topic assigned; 30-36% had two topics and only between 9 and 18% had three topics. In contrast, for even 65-75% of the sentences of the review data set only one topic has been assigned. This reflects the observation reported before that topic detection for reviews is more obvious. An other observation is that the most probable topic has mostly been selected as correct at least for WebMD and the customer review dataset. The topic selection based on LDA assigned probabilities has been proven successful.

Relating the accuracy values to the sentence length (after normalization) shows that even though the sentences are quite short, good results are achieved. The average sentence length after normalization for sentences of the review data set is 3.95 which is even shorter than sentences of the other data sets (WebMD 5.2 words per sentence, Slashdot 4.39 words per sentence). Nevertheless, the best results are achieved for this data set. It seems that the sentence length does not influence the quality of topic detection. More importantly for topic detection and labeling are the words that are used to build the topic models or - at the end - describe the topic.

4.4 Quality of Topics

To measure the suitability of the topic terms to describe a topic, the evaluators were asked in a second evaluation to decide whether the topic terms belong together, i.e. describe the same topic. For the WebMD dataset with $k = 20$ the assigned topic terms are consistent (i.e., zero or one word is not fitting) for 75% of the topics. This is in contrast to other values of k in this dataset, where for 54% (k=100) and 66% (k=50, 150, 200) consistent topic terms are provided.

For the customer reviews, even 84% of the topics can be considered consistent for k=50. For the Slashdot dataset, the topic terms seem to belong to each other to the smallest extent: only 65% of the topics can be considered consistent.

Another aspect of the topic quality is the number of terms that can be found in more than one topic. If for two topics the selected topic terms are overlapping, these topics are more difficulty to distinguish by our test persons. This influences the results of the evaluation presented in the section before, since for topics with the same or similar topic terms it is clearly harder to decide for the correct topic. For example, in the WebMD dataset, around 15% of the topics contain the topic term *time* and 9% contain the term *child*. Other frequently occurring topic terms in this data set are *patient, health* and *people*. For the review dataset, the topic term *player* occurs in average in 21% of the topics. Other terms are *ipod* or *phone*. The largest overlap in topic terms can be identified for the Slashdot topics. The term *people* occurs in 50% of the topics. 10% of the topics contain the term *government*.

Dataset	Sentences	Unique words	Avg. Sentence Length (words)	Optimal k
WebMD	63,169	2754	5.2	20
Customer Reviews	7,757	344	3.95	50
Slashdot	758,504	10,165	4.39	200

Table 3. Size of the three datasets

We can conclude, that the topics identified for the Slashdot dataset are most similar to each other. They can be discriminated from each other less properly due to overlapping topic terms. This certainly also influences the quality of topic detection.

4.5 Characteristics of the Datasets

The topic analysis by exploiting LDA allows to compare the data sets and to characterize them in terms of topics used. Table 3 shows the characteristics of the three datasets in terms of number of sentences and of unique words as determined by the LDA algorithm. As mentioned in the section before, the optimal number of k (number of topics to be distinguished by LDA) varies between the different datasets. A correlation between the optimal number of topics and data set size could not be identified. Even though the review data set comprises only 355 unique words, LDA performed best with 50 topics to be distinguished. In contrast, for WebMD the optimal values of k is 20, but significantly more unique words are used to calculate the topic models for this dataset.

A frequently occurring topic in the customer review data set is *digital cameras*. Others deal with dvd players, ipod, and cell phones. This reflects the content of the data set (see description before). Surprisingly, the Top 5 WebMD topics do not contain any medical terms, but rather general terms (e.g. *child, parent, school*) or health-related terms (e.g. *fat, calorie*). We can conclude that the blog postings discuss rather general issues such as dietary habits and child care. A large amount of blog postings of this dataset are collected from the WebMD blog called *healthy-children* and *safety4kids* which explains the top topics of this dataset. The Slashdot dataset contains political discussions which is also reflected by the Top 5 assigned topics. The topic terms are domain-specific in a sense that they are mainly words from the political domain (e.g., *politician, government, business*).

5 Discussion

In this paper, the accuracy of topic detection with LDA on sentence-level was examined. Due to missing evaluation material and the uniqueness of this evaluation, the results can not be compared to results of other topic detection algorithms. We expect quality improvements when considering synonyms appropriately (*Microsoft* vs. *MS*) and dealing

with writing error (e.g., *pucture* instead of *picture*) through correction or the use of Soundex mechanisms.

The correlation between the number of topics and the properties of the dataset (number of sentences, number of unique words, sentence length) is not linear. The question on selecting the best number of clusters to be distinguished by LDA remains open. Further research is needed to provide an automatic assignment of the optimal number of topics.

Our evaluation relied upon the annotation of two human annotators. Clearly, the quality of annotation depends on the background knowledge of the annotating person. In particular the medical weblogs were difficulty to understand by the annotators due to the medical terminology used within the posts. Our annotators were both not specialized in medicine. Therefore, their judgements regarding the topics differ due to different background knowledge in the areas. Additional reasons for annotator disagreement were inconsistent or overlapping topics. Further, sometimes it is difficult to decide for a sentence topic if the context of the sentence is missing. Nevertheless, we focused on studying the accuracy of topic detection for sentences since in future work, we intend to apply LDA to Twitter messages which are very short and noisy.

Quality improvements of the chosen approach could be achieved for example by considering synonyms or dealing with writing errors and abbreviations. The LDA-determined topic words describe the topics very well for the optimal number of topics. A substantial improvement could be achieved when frequently used words which do not help to distinguish topics are excluded from being selected as topic terms (e.g. Slashdot: people, WebMD: time).

We have chosen LDA in our work since the calculated probabilities are well suited to exclude irrelevant sentences or sentences without clear topic, respectively. This could be confirmed by the evaluation results. Instead of a clustering approach such as LDA, classification based on keywords is a frequently used topic detection method. But, in an online scenario, new topics come up and therefore, specifying classes in advance or establishing lexicon lists is unsuited in our context.

Since LDA does not rely on syntactic or grammatical features, topics can be assigned correctly, even if sentences are ungrammatical. Similarly, to keyword extraction and lexicon lookup approaches to topic detection, LDA can fail when writing errors occur. A clear benefit of LDA is that topics can be assigned correctly even if no topic terms can be matched explicitly, simply by relying upon the context. We conclude that LDA is suited to assign topics to noisy and also to short sentences correctly.

6 Conclusion

In this paper, we study the quality of LDA on sentence level for noisy and short sentences. The experimental evaluation, with three diverse real-world datasets, demonstrates the applicability of the proposed solution, and shows very promising results. In future, we will study the performance of LDA when considering synonyms in the topic identification algorithm. Another extension would be to refine the results of the topic identification by domain-specific knowledge (e.g., using an ontology, or a relevant term dictionary), in order to only consider topics that are related to the domain under consideration.

References

[1] I. Bíró, J. Szabó, and A. A. Benczúr. Latent dirichlet allocation in web spam filtering. In *AIRWeb '08: Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pages 29–32, New York, NY, USA, 2008. ACM.

[2] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty. Latent dirichlet allocation. *JMLR*, 3, 2003.

[3] C. Chemudugunta, P. Smyth, and M. Steyvers. Modeling general and specific aspects of documents with a probabilistic topic model. In *NIPS*, pages 241–48, 2005.

[4] C. Clifton, R. Cooley, and J. Rennie. Topcat: Data mining for topic identification in a text corpus. In *TKDE 16(8)*, pages 949–964, 2004.

[5] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-manning. Domain-specific keyphrase extraction. pages 668–673. Morgan Kaufmann Publishers, 1999.

[6] C. Huang, Y. Tian, Z. Zhou, C. X. Ling, and T. Huang. Keyphrase extraction using semantic networks structure analysis. In *ICDM '06: Proceedings of the Sixth International Conference on Data Mining*, pages 275–284, Washington, DC, USA, 2006. IEEE Computer Society.

[7] J. Li, Q. Fan, and K. Zhang. Keywords extraction based on tf/idf for chinese news document. *Wuhan University Journal of Natural Sciences, 12(7)*, pages 917–921, 2007.

[8] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *WWW*, pages 171–180, New York, NY, USA, 2007. ACM.

[9] V. Stoyanov and C. Cardie. Annotating topics of opinions. In *International Conference on Language Resources and Evaluation*, 2008.

[10] V. Stoyanov and C. Cardie. Topic identification for fine-grained opinion analysis. In *International Conference on Computational Linguistics (Coling 2008)*, pages 817–824, Manchester, UK, August 2008.

[11] C. Wartena and R. Brussee. Topic detection by clustering keywords. In *DEXA*, pages 54–58, 2008.

[12] D. Xing and M. Girolami. Employing latent dirichlet allocation for fraud detection in telecommunications. *Pattern Recognition Letters*, 28(13):1727–1734, October 2007.