Dynamic Topic Detection and Tracking using Non-negative Matrix Factorization

Michael Tannenbaum[a,b] Andrej Fischer[a] Johannes C. Scholtes[b]

[a] *Comma Soft AG, Bonn, Germany*
[b] *Department of Knowledge Engineering, Maastricht University, Netherlands*

Abstract

Most textual data contains a time component that is highly relevant for many economic decision processes. Especially in large document sets and high-frequency document streams it is impossible for humans to gain an overview of the topics in all texts by reading them. Although algorithms such as non-negative matrix factorization (NMF) or latent Dirichlet allocation (LDA) have been applied successfully for static topic extraction in the past years, only few approaches exist for topic analysis on dynamically changing data. We present a hybrid NMF approach combined with hierarchical clustering to detect and track emerging topics in streaming text data. The performance of the algorithm is demonstrated both on simulated and on real-world news text data.

1 Introduction

With the growth of the internet, companies are confronted with a rapidly increasing amount of text data from emails, social media, forums, support tickets, news sources and many more. As a consequence, it is impossible for a single person to keep track of all relevant text data in most cases. Texts need to be read by multiple persons which is expensive, leads to inconsistent results due to different perception of readers and makes it hard to detect global changes in the data.

At the same time, fast interaction with customers through the internet opens up great opportunities: companies have the potential to provide a higher service level by reacting almost immediately to texts created by customers but also to any other text source such as world news. Realizing this potential requires intelligent algorithms which are able to react to emerging topics as fast as possible and at the same time track existing topics over long time spans. Although static topic modeling algorithms such as non-negative matrix factorization (NMF) or latent Dirichlet allocation (LDA) have become quite popular in the last 15 years and standards for the extraction of static topics have evolved, extensions of these algorithms that can deal with dynamic data are still rare.

Therefore we present a novel hybrid approach based on NMF combined with hierarchical clustering. We describe the underlying NMF model and the modifications needed to create a stable dynamic NMF algorithm which is able to detect new topics and, once found, to track topics over time.

2 Related Work

Previous approaches differ widely in terms of their interpretation of the problem, the granularity of detected topics and the algorithms used. First noteworthy approaches have been presented between 1996 and 2004 in the context of a DARPA-sponsored common-task research program in Topic Detection and Tracking (TDT) which interpreted the task as the detection of events in news texts and their tracking among other news texts [1]. Most approaches originating from this program extract named entities to find entities representing events. These named entities are related to other entities or words in the context and tracked among other texts.

Other topic modeling methods used for the extraction of static topics from a predefined set of texts are Probabilistic Latent Semantic Indexing (PLSI) [7], Non-negative Matrix Factorization (NMF) [8] and Latent Dirichlet Allocation (LDA) [3]. The last three algorithms define generative probabilistic models in which each text is modeled as a finite mixture over an underlying set of latent topics where each topic is defined by a characteristic distribution over words. The models are trained by optimizing a certain objective function that measures the quality of the reconstruction of actual term frequencies by the model.

Some variants of these algorithms have been published in the last decade to allow the tracking of topics. Cao et al. [4] and Wang et al. [11] allow the dynamic tracking of changes in topics for streaming data using NMF. They impose similar update rules for the topic model and update the model with the goal to find better representations of new texts appearing from a data stream. Both approaches however assume that the number of topics is fixed and new topics can therefore not be detected.

Other approaches which do include both the detection of new topics and the tracking of existing topics have been published recently. Saha and Sindhwani [9] proposed an algorithm which learns emerging topics from data streams using NMF. Emerging topics are detected by an increase rate of word frequencies which implies that only quickly emerging topics can be found. A similar LDA-based approach has been proposed by AlSumait et al. [2]. In contrast to [9] it assumes each non-matching text to represent a new topic and refines this “topic” when more texts appear.

3 Topic Extraction using NMF

First we describe the generative model that is used to extract topics from a static set of texts using NMF before proceeding with the dynamic approach in section 4. The model is based on the bag-of-words (BoW) model in which every text in a set of documents $\{D_i\}_{i=1..n}$ is represented by the frequency of each word j from the vocabulary $\{\Omega_j\}_{j=1..m}$ in the respective text. The result is a vector $C_{i\cdot}$ for each text i that contains all the word frequencies and sums up to the total number of words in a text N_i. The generative model assumes that every text is composed of a finite number K of topics. This means that the word frequencies of each text – i.e. each vector $C_{i\cdot}$ – can always be explained by a linear combination of these topics.

The affiliation of each text to each topic is expressed as a fraction. According to the assumption that each text only consists of the K topics and no other components, the fractions of all topics for one text must always add up to 1:

$$\forall i : \sum_{k=1}^{K} W_{ik} = 1, \, \forall k : \, W_{ik} \geq 0 \tag{1}$$

These fractions form $W_{i\cdot}$, the topic affiliation vector of text i. Each topic k is in turn defined by a rate for each word in the vocabulary to appear in a text which consists solely of topic k. These rates form the topic signature vector $H_{k\cdot}$. If there is no feature selection, i.e. all words in the vocabulary are considered in the model, the sum of each $H_{k\cdot}$ is also 1:

$$\forall k : \sum_{j \in \Omega} H_{kj} = 1, \, \forall j : \, H_{kj} \geq 0 \tag{2}$$

The probability of any word in a text i consisting of the topics W_i to be word v_j in the vocabulary is therefore:

$$P(v_j | W_{i\cdot}, H) = \sum_{k=1}^{K} W_{ik} H_{kj} \tag{3}$$

The generative model for texts is defined by a multinomial distribution M with the probability given in equation 3 as the probability of each possible outcome. Note that this concept is similar to PLSI and LDA. The word frequencies of each text are generated by drawing N_i words from the vocabulary given these probabilities:

$$C_{i\cdot} \quad \sim \quad \text{Multi}\big(N_i, \, \big(P(v_1|W_{i\cdot}, H) \quad \dots \quad P(v_m|W_{i\cdot}, H)\big)\big) \tag{4}$$

The expected frequency of word j in text i is thus $N_i \sum_k W_{ik} H_{kj}$. If all C_{ij} are assumed to be distributed independently, they follow a Poisson distribution with $\lambda = N_i \sum_k W_{ik} H_{kj}$. The assumption of independence among term frequencies is clearly an approximation, but allows to represent single term frequencies by a probability distribution and is therefore common in probabilistic term frequency models. As an optimization of the Poisson likelihood of the model is expensive, the Poisson distribution can be further approximated by a normal distribution. From the Poisson distribution, the expected value μ and the variance σ^2 are $\mu = \sigma^2 = N_i \sum_k W_{ik} H_{kj}$:

$$C_{ij} \sim \text{Pois}\left(N_i \sum_{k=1}^{K} W_{ik} H_{kj} \right) \approx \mathcal{N}\left(\mu = N_i \sum_{k=1}^{K} W_{ik} H_{kj},\, \sigma^2 = \mu \right) \tag{5}$$

$$\Longleftrightarrow \; N_i \sum_{k=1}^{K} W_{ik} H_{kj} \gg 1 \tag{6}$$

Empirical comparison of the topics which can be extracted using Poisson and normal distribution shows that differences in the resulting topics only occur in details. Note however that the condition imposed by (6) does not hold for most terms: According to Zipf's law [12], most words appear rarely which implies that the expected frequency of most words in most texts is below 1. If the error introduced by this approximation is a concern, the likelihood can be calculated based on the Poisson distribution instead.

Non-negative Matrix Factorization is used to optimize W and H such that the distance d between C and the product of $N \circ W$ and H becomes minimal where $N \circ W$ denotes the row-wise product of N and W. The algorithm iteratively fixes the matrices W and H in turns during the optimization process. The matrix which has not been fixed is optimized with respect to an objective function f which is given by the likelihood function of the above normal distribution. As the objective function is only used to compare results during the optimization, it must only be guaranteed that $\arg\min(f(C, W, H)) = \arg\min(d(C, W, H))$. Consequently, the log-likelihood of the model can be optimized:

$$l\left(\mu, \tilde{\sigma}^2; C\right) = \log\, L\left(\mu, \tilde{\sigma}^2; C\right) = \frac{nm}{2}\log(2\pi\tilde{\sigma}^2) - \frac{1}{2\tilde{\sigma}^2}\sum_{i=1}^{n}\sum_{j=1}^{m}(C_{ij} - \mu_{ij})^2 \tag{7}$$

Note that an equal variance $\tilde{\sigma}^2$ is assumed for all term frequencies in this approximation. As the terms $\frac{nm}{2}\log(2\pi\tilde{\sigma}^2)$ and $\frac{1}{2\tilde{\sigma}^2}$ are constant and $\frac{1}{2\tilde{\sigma}^2}$ is always positive, they have no impact on the optimization. Only the residual term needs to be maximized:

$$\arg\min(f(C, W, H)) = \arg\max(-\sum_{i=1}^{n}\sum_{j=1}^{m}(C_{ij} - \mu_{ij})^2) \tag{8}$$

Instead of maximizing the right hand side of equation 8, its negated value can be minimized. The term equals the definition of the squared Frobenius norm which is defined as the euclidean norm of the flattened matrix [7]. This allows an efficient calculation of the distance between the actual term frequencies and their reconstruction by the model and thus speeds up the optimization of the model.

Instead of using actual term frequencies, it is common to apply Term Frequency-Inverse Document Frequency (TF-IDF) weighting to the term frequencies. TF-IDF weights term frequencies by their inverse document frequencies, i.e. the inverse number of documents each word appears in. It can be used with this approach to assign higher weights to specific words which can be assumed to be better suited for the distinction of topics. This allows the algorithm to find better delimited topics and thus also helps the user understand the topics. However, the resulting algorithm cannot be connected to a probabilistic generative model for texts. As the expected value $N_i \sum_k W_{ik} H_{kj}$ for the frequency of word j in text i is derived from term frequencies, the use of TF-IDF is a heuristic deviation from the statistically justifiable model.

Feature selection can be enhanced within a language domain by linguistic knowledge in the form of part-of-speech (POS) tagging or named entity recognition (NER): Parts of speech such as pronouns or prepositions are usually irrelevant for the definition and distinction of topics. It is therefore reasonable to include only those parts of speech in the bag of words which are considered relevant – a common choice are all nouns. Thinking this further, named entities can be assumed to be almost *always* relevant which makes it worth considering to limit the features of the bag of words only to these terms. This is expected to yield a high precision, however recall is likely to be impaired as not all topics might be represented by named entities. This makes POS tagging preferable to NER for feature selection in most cases.

4 Dynamic NMF Approach

After detecting topics on static texts, we now add a dynamic component to the model, with the aim to detect new topics and to track existing topics in time. The dynamic approach is based on a sliding window of width Δ_t on the time axis. This window contains the Δ_t latest texts from a data stream. All analyses are performed on texts within the window; all other texts are considered as past. The algorithm is robust to the choice of Δ_t if it is chosen larger than the expected maximum size of a batch of documents that may appear on the data stream.

The algorithm is initialized with a *starting set* of texts S. These texts are used to gain information about the domain and to learn an initial set of topics in the texts using the static NMF-based topic extraction. Starting from this point, new data appear in a streaming manner. As soon as a batch of new texts appears on the data stream, topic weights are assigned to these texts according to the established model. The assignment of topic weights to a text yields the distance d_i between the text i and the current model. This measure indicates the ability of the model to reproduce the text's word frequencies. A high distance d_i means that the existing topics are not sufficient to reproduce the term frequencies of text i whereas a low d_i indicates that the text matches the model.

If a text is properly represented by the model, i.e. d_i is below a particular threshold, there is no need to change the model. The topic weights which have been assigned to the text are thus added to the W matrix. If d_i is high, this means that the term frequencies of text i cannot be expressed as a linear combination of the rows of H. These texts are appended to a new set called the *emerging set* E and are not added to the model for the time being. The emerging set is subject to the bounds of the sliding window and serves as a pool for texts which currently cannot be assigned to the model because of its inability to represent them properly. They are candidates for texts which may change the current model. The concept of an emerging set – even though used differently – was first introduced by Saha and Sindhwani [9].

If the distance of new texts from the model is comparable to the distance of texts in the starting set, they should be considered as matching. The distance of a new text is thus compared with the third quartile $Q_{0.75}$ of the distances of all texts in the starting set S. The percentile rank can be adjusted to make the algorithm more or less strict to incoming texts.

4.1 Detection of new Topics

Texts matching the model are unlikely to contain new topics. Only the texts in the emerging set are therefore considered as candidates for new topics. Given that the model was unable to represent these texts properly, it is reasonable to expect that the texts in the emerging set contain information which is not contained in the current model. Emerging topics are thus detected by finding groups of similar texts within E: If there are several texts which do not match the established model but have a common structure, they are assumed to form a new topic. These common structures are found by clustering the texts in the emerging set. The details of the clustering are described in section 4.1.1.

It is important to note that the common structures resulting from clustering should be different from structures which can already be represented by the model to avoid reproducing existing topics. It is thus necessary to ensure that topics are only created from novel structures. For any term frequency vector there is always a vector $\vec{v}$ which could be added to the linear combination of the topic fractions in W and the word rates in H and would allow a perfect representation of the text by this linear combination. Therefore, we propose the following approach to detect potentially new topics:

$$C_i = N_i \sum_k W_{ik} H_k + \beta \vec{v}_i \tag{9}$$

The variable β denotes the (undefined) coefficient of the new vector for the sake of completeness. This new vector $\vec{v}_i$ could be added to H which would enable the model to represent the new text thoroughly with β as the K^{th} value of the corresponding new row of W. If this was done for all texts which do not match the model however, H would grow rapidly and the new rows of H would hardly qualify as "topics" because each would only represent the error vector of one text. The $\vec{v}$ vectors thus need to be stored in the emerging set until several texts show similar feature patterns. These similar feature patterns are found by clustering the $\vec{v}$ vectors of all texts in the emerging set.

4.1.1 Clustering of the Emerging Set

The goal of the clustering is to find similarities in the term frequencies of texts which did not match the model. As all texts in the emerging set may contain the same new topic, separability of clusters is not relevant. One requirement to the clustering algorithm is thus that it should find clusters with high intra-cluster similarity, whereas algorithms that maximize the inter-cluster distance are unsuitable. Moreover, the number of clusters cannot be known upfront. The task is solved by agglomerative clustering with average-linkage for cluster distances. This clustering however is expensive and hardly scalable: Its complexity is $O(n^3)$ with the number of documents [6]. The number of texts in the emerging set must therefore strictly be limited which is guaranteed by the fixed size of the sliding window.

When a cluster of documents has been found, it is to be decided if it is suited to form a new topic. This step is crucial as both false positives and false negatives can lastingly impair the quality of the model. There are three factors which need to be considered for this threshold:

1. **Intra-cluster distance:** The intra-cluster distance d_I indicates the similarity of word frequencies in the documents of a cluster. A low value accounts for more specific topics.
2. **Support:** A new topic should only be created if there is sufficient support, i.e. if the topic appears frequently in the emerging set. This is necessary to avoid niche topics which only appear in a small number of texts. The support s is measured by the number of texts in a cluster.
3. **Difference from existing topics:** There is a risk of existing topics being detected by the clustering. The minimum distance Δ_γ of an emerging topic γ from all existing topics thus needs to be considered in the decision whether a cluster is suited to represent a new topic.

For reasons of performance, two thresholds are defined: The topic emergence threshold T_E which is calculated from the factors which are given directly by the cluster (factors 1 and 2), and the topic distance threshold T_Δ which defines a limit for the distance from all other topics. All three factors are trained based on the granularity of the existing topics and require information about an exemplary cluster that is suited to represent a topic. Clusters are therefore formed for all topics in the starting set: For each topic k, all texts with a strong affiliation to this topic are drawn from the starting set. Text i is considered to have a strong affiliation to topic k if the affiliation score is more than twice the expected value in a uniform distribution of topics. The word frequencies of all texts that satisfy this condition are clustered hierarchically. Factors 1 and 2 from above are learned from the resulting cluster containing all texts satisfying the condition and are given by the cluster directly.

The intra-cluster distance d_k of each topic k is calculated for all topics and the average of the results determines the threshold T_E for the emergence of new topics:

$$T_E = \frac{\sum_k d_k}{K_{t_o}} \tag{10}$$

The same is calculated for each topic candidate γ that results from the clustering. A cluster emerging at time point t must consequently satisfy the following condition to form a new topic:

$$d_\gamma \overset{!}{\leq} T_E \tag{11}$$

If the above condition is satisfied, the minimum distance of the cluster from all other clusters δ_γ is compared with the minimum distance between the clusters of the starting set δ_k. The minimum distance from all other topics is relevant in both cases because a new topic should not be similar to any existing topic. δ_γ and δ_k are therefore:

$$\delta_\gamma = \min\ \{\|\gamma - W_k\|\}_{k \in K_{t_0}}\ ,\ \delta_k = \min\ \left\{\left\|W_k - W_{\tilde{k}}\right\|\right\}_{\tilde{k} \in K_{t_0} \setminus k} \tag{12}$$

The second threshold is defined by the condition:

$$\delta_\gamma \overset{!}{>} \frac{\sum_k \delta_k}{K_{t_0}} \tag{13}$$

If conditions 11 and 13 hold, a new topic is added to the model. This means that a column is added to W and a row is added to H. This row is initialized by the average vector of all members from the cluster γ. The model is subsequently updated using NMF.

4.2 Topic Tracking

Finally, we need to include the ability to track existing topics. Topic Tracking includes the assignment of existing topics to new texts and the update of existing topics. The first is equal to the update of a row of W by NMF. The second means that the H matrix must be updated according to word frequencies in new texts. If H is updated using only the latest texts however, there is a chance of overfitting to these texts. To gain more stability, H is therefore updated based on *all* previous texts. Old topics thereby gain stability because the new texts only represent a small fraction of the texts that influence the update of H whereas young topics which may not have been defined perfectly at the time of their emergence are refined by additional texts.

Whenever a topic becomes irrelevant, it is deleted from the model as it unnecessarily adds complexity to the optimization. The relevance of a topic is determined based on the topic assignments of all texts in the sliding window: If there is no prior to the distribution of topic weights, the expected affiliation of any text to any topic is K^{-1}. If the average affiliation of all texts within the sliding window to one topic is below $\varepsilon \cdot K^{-1}$ with $\varepsilon \ll 1$, this topic is assumed to be irrelevant and is removed from the model.

As words which were not included in the vocabulary or seemed unimportant before tend to be crucial representatives of changes in the data, it is important to identify these words and add them to the vocabulary. Words which become important for a new topic are assumed to show a lift in frequency. New words are therefore added to the vocabulary if their average frequency in the sliding window is high and significantly higher than in the starting set.

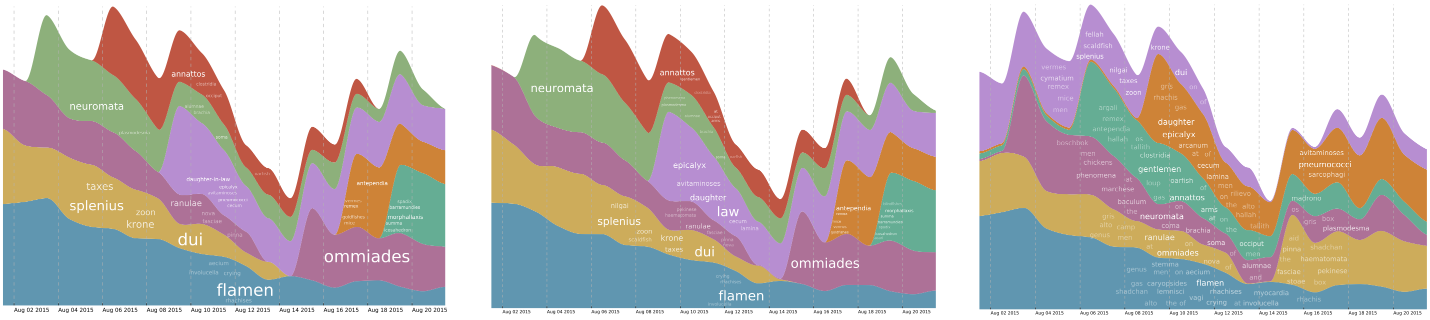
5 Evaluation

In order to be able to test the model without influences from discrepancies between real world data and the model, 100 dynamically changing data stream instances – defined by matrices W and H for each time point – were generated and used to create texts according to the generative model with an average of 1000 words per text, 5 randomly chosen words as topic-specific keywords for each topic and all other words in the vocabulary being randomly added to all texts. Figure 1 shows three exemplary stacked graphs of which the first directly represents the gold standard and the second was learned from the generated texts with a window size of one day. Each learned topic was matched to the most similar topic in the gold standard if 3 out of the 5 most relevant words describing each topic matched. Precision was measured by the fraction of topics which could be matched to the gold standard. If a topic was detected twice, only the first was counted as a correct match. Recall was measured by the fraction of topics in the gold standard which were detected correctly. The results in table 1 demonstrate that texts which closely follow the model can be reconstructed with a high accuracy.

Figure 1c has been created by extracting topics from all texts using pure NMF and splitting the resulting topic assignments into time intervals for comparison. It is missing the two last appearing topics

Precision	Recall	F_1 score
0.954	0.981	0.967

Table 1: Evaluation results using simulated data



(a) Development of topics in the gold standard

(b) Development of topics as reconstructed by the model

(c) Pure NMF reconstruction

Figure 1: Reconstruction of simulated data

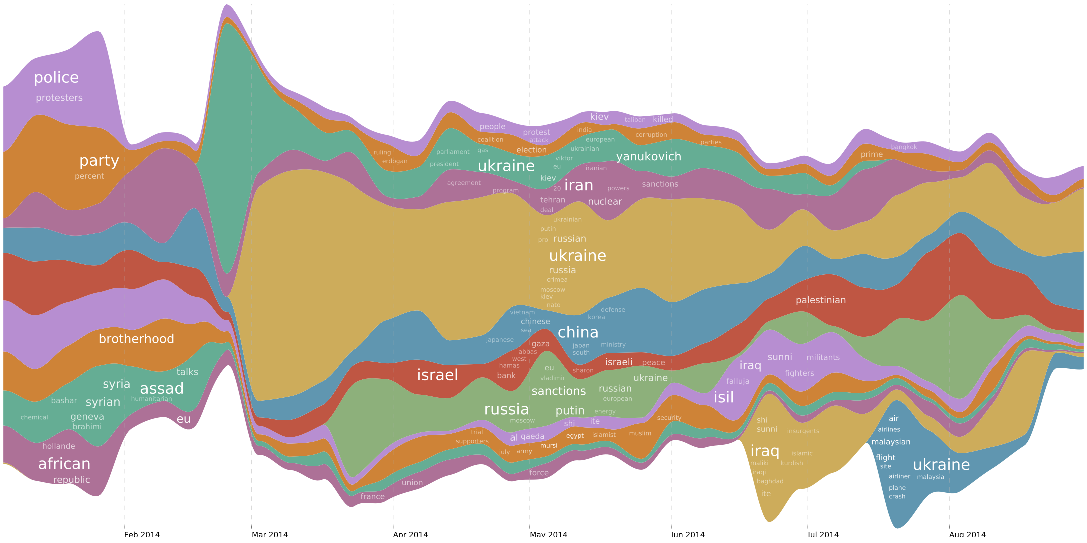


Figure 2: Topic stream visualization of the results obtained from Reuters news data [10]. Data taken from the *World* category, published between January and August 2014.

	Fraction of positive votes			Cohen's κ		
	Unweighted	TF-IDF	POS	Unweighted	TF-IDF	POS
Question 1	0.765	0.843	0.789	0.542	0.557	0.630
Question 2	0.745	0.733	0.684	0.418	0.412	0.607

Table 2: Voting results

~~as they~~ are relevant for their respective time intervals, but not relevant enough in the entire period.

Figure 2 shows the development of topics, learned with a window size of 600 documents, in a set of 15889 texts from the category *World* which have been scraped from the Reuters news archive [10] from January until August 2014. The words in each stream represent the most relevant words describing each topic. The height of each stream is determined by the sum of the respective column of W and thus shows the importance of the topic at each time point. The graph clearly shows the emergence of topics describing the Ukraine crisis, starting with the deposition of Viktor Yanukovych on September 22, a topic about the Crimea conflict emerging one week later, a topic about sanctions against Russia and vice-versa in the middle of march and a topic about the Malaysia Airlines flight MH17 which was shot above Ukraine on July 17. This evaluation also shows that a window size of 600 documents was already sufficient to solve a complex problem. Larger window sizes did not lead to significant improvements.

Equivalent streamgraphs, learned from the news texts in the year 2015 (until July 20) have been handed to ten judges. The judges were asked to evaluate whether the words in each stream describe a common topic (question 1) and whether the words can be related to a real topic in the news (question 2). One streamgraph was learned without term frequency weighting, one with TF-IDF weighting and one with TF-IDF and additional feature selection using POS tags, using only nouns and verbs. The middle column of table 2 shows the average fraction of positive votes in the tests. The second question is intuitively the more complex one as it requires additional knowledge about the main topics in the news which leads to less positive answers to this question. The table shows that the fraction of positive answers to the first question obtained using TF-IDF weighted term frequencies is considerably above the one obtained without term frequency weighting and also above the one obtained using POS tagging according to the judges.

In order to assess the relevance of the results, the inter-judge agreement was measured using the κ measure proposed by Cohen [5]. The pairwise κ values have been averaged in order to obtain the agreement among all judges. Table 2 shows that there is a reasonable agreement among judges and that κ values are similar between the results without POS tagging. The highest agreement could be reached with POS-based feature selection. The evaluation shows that both approaches – according to the judges – lead to meaningful results. Best results from a human perspective could be obtained using TF-IDF weighted term frequencies.

6 Conclusion

A novel approach for the detection and tracking of topics in streaming text data has been elaborated and presented. A dynamic topic model has been developed based on non-negative matrix factorization. It extends the static approach by the ability to handle changing data and allows the model to adapt to these changes. This is accomplished by using hierarchical clustering to find similarities in texts which cannot be explained by the model. The approach has been demonstrated to produce promising results both on simulated and on real world data. It is applicable independently of the domain, the granularity of topics and even of the language if no linguistic features are used.

In order to allow the comparison of different approaches, a corpus providing a ground truth is needed. The goal of the DARPA TDT competitions was to find story boundaries within texts and to map stories to topics. The DARPA TDT corpora therefore provide a ground truth for these tasks [1]. Our approach however operates on whole texts and cannot provide story segmentation. This makes approaches hardly comparable and makes a ground truth for probabilistic term frequency based approaches desirable.

There are several prospects for future work. As the clustering is not incremental, all texts in the emerging set are clustered each time new documents are added. Runtime could thus be improved by incremental clustering. Feature selection could further be improved using linguistic knowledge: Many words appear in different word forms. It is thus reasonable to use stemming in order to reduce the number of features which need to be maintained and improve the choice of words which are presented to the user.

References

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, Lansdowne, VA, February 1998.

[2] L. AlSumait, D. Barbara, and C. Domeniconi. On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on*, pages 3–12, Dec 2008.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *The Journal of machine Learning research*, 3:993–1022, 2003.

[4] Bin Cao, Dou Shen, Jian-Tao Sun, Xuanhui Wang, Qiang Yang, and Zheng Chen. Detect and track latent factors with online nonnegative matrix factorization. In *IJCAI*, pages 2689–2694, 2007.

[5] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[6] William H.E. Day and Herbert Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, 1(1):7–24, 1984.

[7] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

[8] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, October 1999.

[9] Ankan Saha and Vikas Sindhwani. Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 693–702. ACM, 2012.

[10] Thomson Reuters. Reuters site archive. Retrieved from `http://www.reuters.com/resources/archive/us/`, July 2015.

[11] Fei Wang, Chenhao Tan, Ping Li, and Arnd Christian König. Efficient document clustering via online nonnegative matrix factorizations. In *Eleventh SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, April 2011.

[12] George Kingsley Zipf. *Human Behavior and the Principle of Least-Effort*. Addison-Wesley, Cambridge, MA, 1949.