InstaVis: Visualizing Clusters of Instagram Message Feeds

Andreas Stöckl*, Jeremiah Diephuis†, and Andrea Aschauer‡
Digital Media Department, University of Applied Sciences Upper Austria
Hagenberg, Austria
Email: *andreas.stoeckl@fh-hagenberg.at, †jeremiah.diephuis@fh-hagenberg.at, ‡andrea.aschauer@fh-hagenberg.at

***Abstract*—We provide a method for visualizing the information associated with the clusters used for topic modeling of *Instagram* Message Feeds. For this purpose, a series of interactive dashboards are used to determine the right number of clusters and a suitable interpretation of each cluster. These extend previous approaches for regular text documents and focus on including specific information in *Instagram* feeds such as hashtags and linking structure.**

***Index Terms*—visualization, data clustering, interaction**

I. Introduction

Due to the continually increasing number of written texts, machine classification is becoming more and more important. Machine learning and clusters enable a preselection of texts and assign them an additional objective level of meaning. This objective "machine view" can facilitate the identification of new and potentially valuable relationships, but the additional cluster information also requires human interpretation. For most people, high dimensional vectors are very abstract constructs. In order to make this data human-interpretable, different perspectives on the data are required.

The use of topic modeling with the LDA method [1] inevitably leads to questions about the interpretation of the method's results (referred to as the "model checking problem" by [2]), which are probability distributions over the words in the corpus. For example, a list of the most frequent 30 words per topic including frequency distribution can be used to interpret the content of the cluster. However, this is only a first approach to solving the problem, and interactive visualizations are often employed to gain better insights.

Many of the available approaches are only partially applicable to the analysis of *Instagram* feeds because, on the one hand, these feeds often contain limited textual information, but on the other hand, they contain additional information through the use of hashtags and links to other feeds. These can be especially useful for clustering and interpretation purposes. We have adapted and extended existing methods for the visualization of topic modeling with LDA to meet these special requirements.

II. Related Work

A number of visualization approaches have been proposed for representing the content of larger sets of text documents. *Termite* [3], for example, is a system that uses heatmaps to compare topic-word distributions in text documents corpus-wide, but without any interactivity. *TopicNets* [4] is a web-based system for visual and interactive analysis of large sets of text documents using topic models. It presents a range of visualization types and interaction mechanisms and uses dimensionality reduction to plot documents in a 2D space, but does not show topic or document composition. The *Topic Navigator* of [5] also enables document interactivity but does not show comparative topic distribution among documents.

LDAvis [6] offers a particularly interesting approach that on the one hand shows the distance of the clusters from each other by means of projection in 2D, but also shows the word distributions of the selected clusters via click. The sorting order of the word distributions can be influenced by setting a parameter.

Also of interest is *Topic Explorer* [8], which displays the topic distribution within each article, in addition to the weight of that topic in the article. Hovering over a topic shows the top ten words in that topic and highlights the distribution of that topic across selected documents.

Despite the general usefulness of these approaches, they still lack a higher degree of flexibility in providing different views of complex data sets such as *Instagram* feeds. A more modular and interactive approach would be advantageous, which will be demonstrated in the visualization approach we will refer to as *InstaVis*.

III. Data and Model

To demonstrate this approach, a collection of available *Instagram* posts were downloaded for local analysis (using the service *Storyclash* [1]) and the text per post present is stored in each feed. The hashtags and links to other feeds are extracted and stored separately. Then, for each feed, the text, hashtags, and links are combined into one document at a time.

For data cleansing and pre-processing of the vocabulary, a part-of-speech tagging is performed using the NLP package *Spacy* [2]. The package is used to break the documents down into individual tokens and to filter out punctuation, spaces, and some word types. In addition, the words are brought back to the root of the word by means of lemmatization. Before the Topic Model is calculated, tokens can be filtered out that

[1] https://www.storyclash.com/
[2] https://spacy.io

978-1-7281-9134-8/20/$31.00 ©2020 IEEE
DOI 10.1109/IV51561.2020.00076

have a certain total number of occurrences in all documents or occur in more than a certain percentage of all documents. If desired, certain individual words can also be additionally excluded. The LDA model is then calculated with the package *Gensim* [9], indicating the number of clusters to be calculated.

IV. Visualization via Dashboards

The dashboards were created with *Plotly Dash* [3]. As in the original *LDAvis* module, we use TSNE [7] to calculate and display projections of the documents and clusters in 2D. As an extension to *LDAvis*, our dashboard (Fig. 1) can display not only the list of the top 30 tokens for the selected clusters, but also the top 30 *Instagram* feeds of the cluster ordered by the share of the selected topic in the feed.

Another dashboard (Fig. 2) displays the projections of the individual feeds including the cluster membership according to the selected color scheme. In this display, feeds can be selected either by using tools such as the lasso tool or by selecting specific feeds in a dropdown field and defining a distance radius. For the selected feeds, the most common words can now be displayed, including hashtags and links to other feed on the right side of the dashboard, sorted by frequency.

In a third dashboard (Fig. 5), the linking information between the feeds is displayed directly in the projection of the feeds, with corresponding lines between the points of the feeds. By means of a parameter, the number of mentions of a feed from which this line should be displayed can be set. In this way, a clear overview can be created for many connections by hiding connections that are only rarely used. This dashboard can be used to analyze whether the topic structure provided by the LDA procedure matches the linking structure. Since it can be assumed that feeds mainly link other feeds from the same topic, links in the displayed graph should be mainly within the clusters (displayed with the same color code).

V. Example of an analysis with InstaVis

InstaVis consists of a sequence of interactive graphical dashboards, input options and data tables. We start with a selection of 40,000 *Instagram* messages collected by *Storyclash*. Here we could filter for posts from special accounts.

In the next step, we merge the texts, hashtags and mentions of all accounts into one document. Here we can set the minimum number of posts that must be present for an account to be included in the further analysis. In this example, we select 10 to ensure that at least a minimum amount of text is available for the analysis with LDA. For each account, we also have the human-made assignment to a category in this data set, which we will use later for evaluation purposes.

In the provided example, 654 *Instagram* accounts are thus available in the form of the texts of at least 10 contributions each (text, hashtags, links). For these, a vocabulary and a text corpus is now formed. Certain word types, which were determined by a part-of-speech tagging, are filtered and the minimum word length can also be specified.

[3]https://plot.ly/dash/

Afterwards, manually selected words can be additionally filtered. In the example, we filter “bio” and “link”, as they occur in a large number of posts without information content for analysis. Now we can start to find the topics in the accounts by calculating an LDA model, starting with six clusters. Here we reduce the vocabulary again to keep the dimension of the vector space within limits by filtering rare words that occur fewer than five times and words that are very general and occur in over 50 percent of all accounts. These two limits are adjustable.

To be able to visualize the high-dimensional vector representations of the accounts (counts of all words in the vocabulary), we need a projection in 2D and employ TSNE for the following visualizations; this offers the best results in practice. First we consider the projection of the cluster centers into the plane (analogous to the representation in *LDAvis* [6]) together with the distribution of the most common 30 words (Fig. 1). By selecting a cluster center point, the word distribution can be restricted to this cluster. The size of the circles indicates the size of the cluster in terms of the number of accounts belonging to the cluster. Thus, in this diagram it is visible that Topic 5 contains only very few accounts, and a closer examination or reduction of the number of clusters should be considered.

In Fig. 1, the cluster with the number 4 (yellow) is selected. The most common 30 words of the cluster are also highlighted in yellow in the histogram on the right. This shows both the absolute frequency and the relative share by comparing them with the shares in the other clusters indicated by the corresponding color. An examination of the top words: “star, music, getty, song, album, etc.” facilitates an interpretation of the topic. The word “getty” would be a candidate for filtering, since it is most likely part of the image rights label “getty Images”. By changing the “Lambda” parameter, the order of the top words can be changed, depending on how strongly the absolute part or the specific part in the cluster should be emphasized. For a more detailed description of the parameter, see [6].

Another way to interpret the clusters is not to look at the most common words, but to look at the titles of the feeds with the highest proportion of this topic. For Topic 1, this is shown in Fig. 3. In the example, feeds such as “Bernie Sanders”, “Breitbart”, “Michael Moore” and others contain almost exclusively content of this topic, which also permits conclusions about the topic.

However, this type of display has the disadvantage that only the cluster centers and the size of the cluster are shown, making it difficult to see how the distribution of the individual accounts and their distances are. Are the clusters “mixed” or clearly separated? Are individual accounts “outliers”?

Fig. 2 shows a projection representation of the individual feeds including cluster representation by means of color code. Individual feeds can also be assigned to the topics using the color. In the display, individual feeds or groups of feeds can be selected by means of a lasso or rectangle tool. The right side of the dashboard then shows detailed information about

Fig. 1. Cluster distance and word distribution: the absolute frequency and relative share of topics and words can be clearly visualized.

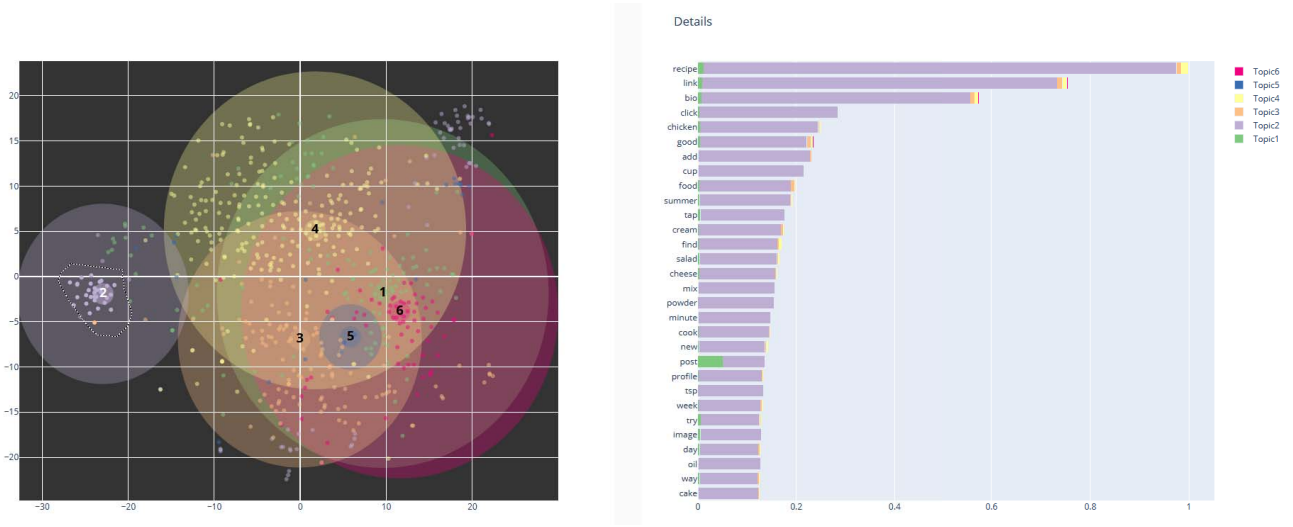

Fig. 2. Cluster and single feeds: Individual feeds can be selected from a projection representation (left) to show a detailed view (right).

the selected topics.

The detailed information can again be the names of the selected feeds, ordered by cluster dominance of the different clusters, or the frequencies of the most important words with percentages in the individual clusters (Fig. 2). The figure clearly shows that the group of selected feeds can be assigned to the food topics based on the word frequencies.

In the detailed information, however, additional information can be selected here. The distribution of the hashtags can be displayed (Fig. 4). As a vocabulary specially chosen by the authors to describe the content, these have a highly informative value, but are not always available in all posts. The example in Fig. 4 confirms the interpretation of the group as a “Food Profile” on the basis of the hashtags. Furthermore, a link list to other feeds can be displayed in the detailed view. This enables the cluster to be interpreted based on the feeds it refers to.

Entire clusters can also be shown and hidden. Individual feeds can be selected by name, and then similar feeds can be selected by specifying a radius. An additional dashboard (Fig. 5) is also available to visualize the links between the feeds combined with the display of the individual feeds including clusters.

This shows whether the linking structure confirms the cluster structure by links within the clusters, or whether there are many links between the clusters. The dashboard offers the possibility to set via a parameter at which number of

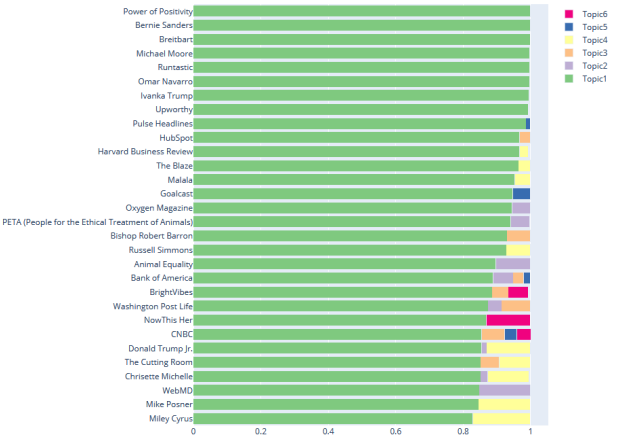

Fig. 3. Cluster distance and document distribution: the titles of individual feeds can be utilized to examine document distribution within categories.

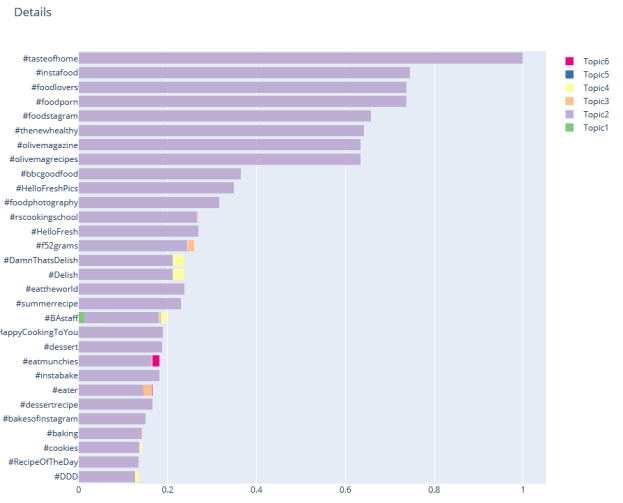

Fig. 4. Distribution of hashtags: although potentially very useful, hashtags are not available in all posts.

occurrences a link is displayed. For example, it can be defined that only links that appear two or more times in the data are displayed.

Finally, it is possible to compare the topics from the cluster analysis with categories assigned by humans. Fig. 6 shows which clusters are composed of which categories and vice versa, and as a further dimension, a list of words can be defined and their share in the clusters visualized.

The categories assigned by people to the feeds are very uneven in size, as can be clearly seen in the example of the category “Media” in Fig. 6, and on the other hand, very different in terms of granularity. Accordingly, for example, all clusters contribute to the category “Media”. The allocation of the clusters found by the LDA procedure therefore does not fit well with the categories assigned by humans.

VI. Conclusion

The visualization shows that cluster interpretation considers many dimensions and human semantics. A combination of tools and information levels is therefore required in order to be able to extract added value from machine-generated clusters. Existing cluster interpretation approaches mostly rely on a single form of visualization.

To improve on this, *InstaVis* utilizes the interplay of various interactive visualizations in the form of dashboards. Preliminary analysis with data from *Instagram* message feeds indicates that more semantic information can be interpreted from the clusters with this approach. Clusters are complex multidimensional data; *InstaVis* demonstrates that a mix of visualization forms can effectively be used to interpret the data from different perspectives.

For the future, it would be interesting to apply these methods to other data sources, such as *Twitter* posts, and compare the results. In such a use case, data structures comparable with hashtags and mentions are also available. However, the different use of text in *Instagram*, which is very much based on visual material, also suggests that there could be differences in the results.

References

[1] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. “Latent dirichlet allocation.” Journal of machine Learning research 3, no. Jan (2003): 993-1022.

[2] Blei, David M. “Probabilistic topic models.” Communications of the ACM 55, no. 4 (2012): 77-84.

[3] Chuang, Jason, Christopher D. Manning, and Jeffrey Heer. “Termite: Visualization techniques for assessing textual topic models.” In Proceedings of the international working conference on advanced visual interfaces, pp. 74-77. 2012.

[4] Gretarsson, Brynjar, John O’donovan, Svetlin Bostandjiev, Tobias Höllerer, Arthur Asuncion, David Newman, and Padhraic Smyth. “Topicnets: Visual analysis of large text corpora with topic modeling.” ACM Transactions on Intelligent Systems and Technology (TIST) 3, no. 2 (2012): 1-26.

[5] Chaney, Allison June-Barlow, and David M. Blei. “Visualizing topic models.” In Sixth international AAAI conference on weblogs and social media. 2012.

[6] Sievert, Carson, and Kenneth Shirley. “LDAvis: A method for visualizing and interpreting topics.” In Proceedings of the workshop on interactive language learning, visualization, and interfaces, pp. 63-70. 2014.

[7] Maaten, Laurens van der, and Geoffrey Hinton. “Visualizing data using t-SNE.” Journal of machine learning research 9, no. Nov (2008): 2579-2605.

[8] Murdock, Jaimie, and Colin Allen. “Visualization techniques for topic model checking.” In Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015.

[9] Řehůřek, Radim, and Petr Sojka. “Gensim—statistical semantics in python.” Retrieved from genism. org (2011).

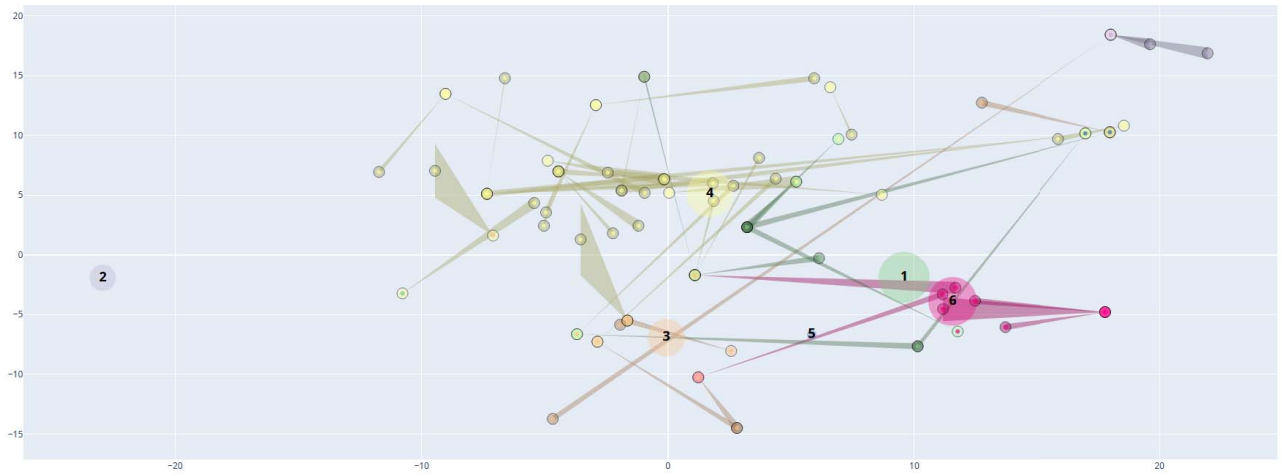

Fig. 5. Cluster with graph of connections: links between individual feeds can be displayed and similar feeds can be selected by defining a radius.

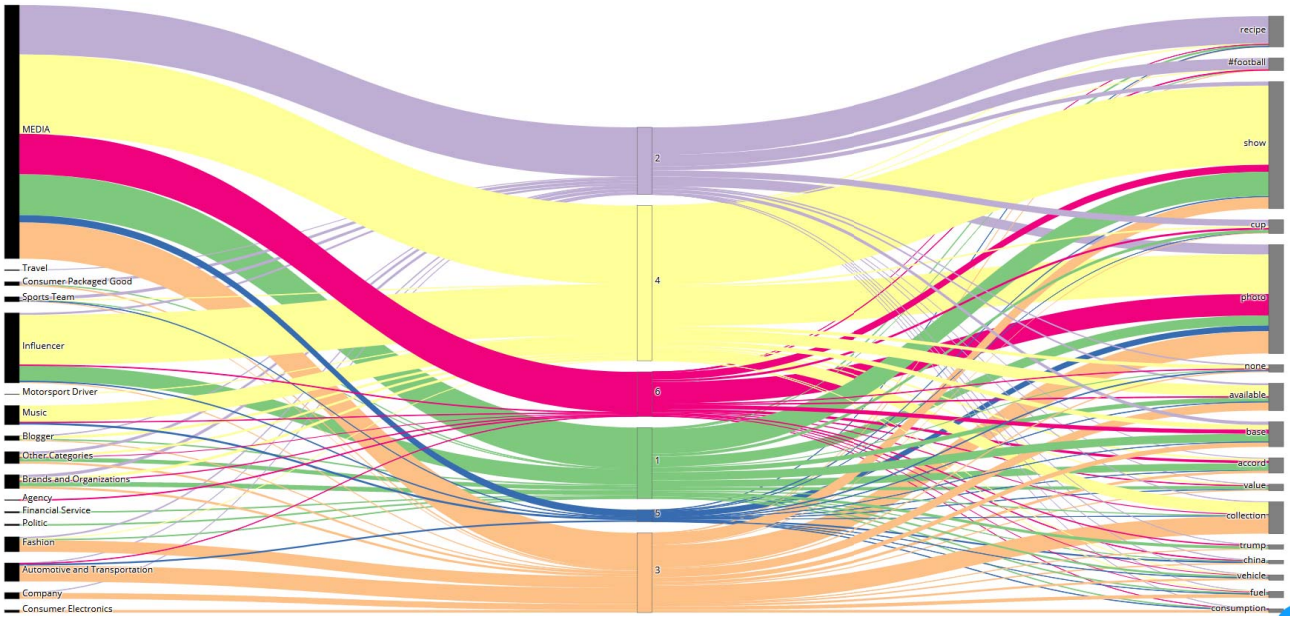

Fig. 6. Clusters with categories and top words: the human-assigned categories (left) are more uneven than those assigned by LDA (right).