Text Insights: Natural Language Analytics for Understanding Social Media Engagement

Frank Grimm, Matthias Hartung, and Philipp Cimiano

Cognitive Interaction Technology Center of Excellence (CIT-EC)
Bielefeld University
33615 Bielefeld, Germany
fgrimm@techfak.uni-bielefeld.de, {mhartung, cimiano}@cit-ec.uni-bielefeld.de

Abstract. We present *Text Insights*, an application for understanding factors of user engagement in Facebook pages. Providing analytics based on natural language processing, Text Insights is complementary to existing tools offering mainly numerical indicators of user engagement. Our system extracts keyphrases from page content in a linguistically motivated manner. Keyphrases are weighted according to their relevance as approximations of the most important topics in the community. We demonstrate that the system provides valuable insights for page owners interested in trend discovery, content evaluation and content planning.

Keywords: social media, text analytics, natural language processing

1 Introduction

Modern companies use their presence on social media platforms for diverse business goals. Social media present a new and unique way for direct interaction between the company and different stakeholders, right down to the customer.

While most social media platforms offer some way to measure user engagement, many focus on customer conversion, rather than content. Tools like *Insights* for Facebook Pages[1] or *Social Analytics* for Google Analytics[2] provide convenient indices to track user demographics and engagement in numerical terms (e.g. page impressions, number of likes or referrals). Most traditional social media metrics rely on a large number of interactions to generate actionable and meaningful insights. While some brands promote themselves through viral campaigns, most efforts in social media base on long-term campaign strategies, rather than short-term or viral approaches[2].

We argue that for evaluating and refining such strategies, it might be beneficial to analyze the textual content of user contributions. Therefore, our *Text Insights* web application offers analytics for investigating the publicly available data on a specific Facebook page on both the numerical and the textual level.

[1] `https://www.facebook.com/insights/`
[2] `http://www.google.com/analytics/`

2 Text Insights

Text Insights analyzes data on a social media presence in the Facebook ecosystem. A Facebook page contains different forms of content contributions by the maintainer of a page and outside commenters. In the following, we outline how this data is acquired (Section 2.1) and presented to the user in an aggregated form (Section 2.2), based on linguistic analysis (Section 2.3).

2.1 Data Acquisition Methods

During **page data retrieval**, the system recursively queries the Facebook Graph API[3] for information stored on a Facebook page, comprising all posts and comments (incl. metrics such as count, creation timestamp, etc.). We also retrieve limited **user data** on all contributors associated with the content, which is anonymized for privacy reasons. The system only retains public data available from most user profiles, such as gender, ISO language and country code.

2.2 Data Preparation

Text Insights aggregates general metadata on posts, comments, user demographics and post types (status, photo, question, link and video), as well as textual information derived from page content.

The system presents the **most important topics** within all posts and comments on the page, condensed into a tag cloud. Each topic is denoted as a single keyword or a keyphrase (i.e., a sequence of keywords; see Fig. 1, left). The size of a tag corresponds to its relative importance which is determined by a pipeline including several steps of linguistic analysis as described in Section 2.3. Colors are used to indicate the source of topics: Topics triggered by the page owner that have not been picked up by the users are presented in light blue, topics triggered by the page and picked up by other contributors in dark blue, topics that have been independently triggered by others in brown.

All data is available in an all-time overview, monthly breakdowns, as well as specific queries for a user-defined time frame. The user interface is interactive, thus enabling, for example, to search for contributions containing a certain keyphrase, navigating to the original contribution(s) mentioning a particular keyphrase, or investigating related keyphrases.

2.3 Linguistic Analysis

Text Insights generates keyphrases from each contribution on the page using the following steps of linguistic analysis:

1. **Tokenization.** After sentence splitting, a *WhitespaceTokenizer* is applied to extract sequences of individual tokens from each posting or comment. Both steps make use of NLTK[4]. All tokens are normalized to lowercase characters.

[3] `https://developers.facebook.com/docs/graph-api`
[4] Natural Language Toolkit `http://www.nltk.org/`

2. **Part-of-Speech Tagging.** NLTK and Ark-Tweet-NLP [3][5] part-of-speech taggers are used to assign word classes to each token.
3. **Keyphrase Extraction and Normalization.** The tagged words of a contribution are searched for linguistically meaningful patterns of words (e.g., compound noun-phrases, verb-noun constructions). Linguistic patterns of different type and length can be configured. Keyphrases containing special characters, stop words, numbers or parts of URLs are rejected. All tokens are normalized using the Lancaster stemmer[6] as implemented in NLTK.
4. **Keyphrase Weighting.** Each extracted keyphrase is assigned a TF/IDF [4] relevance score as given in equation (1), where p refers to a keyphrase, d to a specific post or comment within the set D of all contributions on the page, and $f(p, d)$ denotes the frequency of p in d.

$$\text{tfidf}(p, d, D) = \log(f(p, d) + 1) \cdot \log \frac{N}{|\{p \in D : t \in p\}|} \tag{1}$$

3 Use Cases

Text Insights addresses several use cases, among them *trend discovery*, *content evaluation* and *content planning*. The following examples are the results of applying Text Insights to a Facebook page targeting health care professionals in the domain of hypertension treatment, the "Hypertension Hub"[7] (HH).

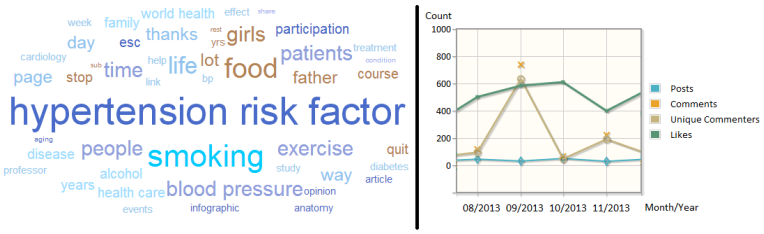

Fig. 1. Tag cloud of weighted keyphrases based on full page content (left); page metrics excerpt for the time interval 08/2013–11/2013 (right)

Trend Discovery. The global tag cloud in Fig. 1 (left) displays specific hypertension risk factors as an overall prominent topic of interest for HH users. Apart from discovering such overall trends, monthly breakdowns can be used for tracking interesting engagement patterns. The unique comment peak for 09/2013 (see Fig. 1, right), for example, can be attributed to contributions surrounding a medical congress ("esc"). The content coverage for this was perceived so well that the congress is still showing in the global tag cloud. As an actionable result, it might be useful to cover similar events in the future.

[5] `http://www.ark.cs.cmu.edu/TweetNLP/`, version 0.3.2.
[6] `http://www.nltk.org/api/nltk.stem.html`
[7] `https://www.facebook.com/thehypertensionhub`

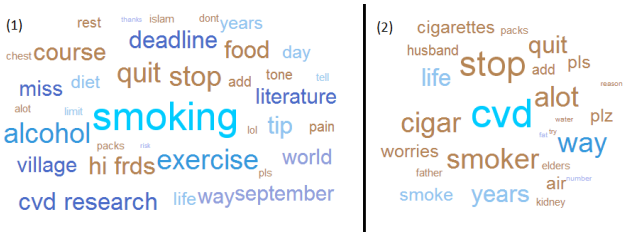

Fig. 2. Monthly keyphrases for 09/2013 (left); keyphrases related to "smoking" (right)

Content Evaluation and Planning. Monthly tag clouds and the ability to gather related keywords to a query term enable the user to evaluate existing content and reactions to the page owner's content. The monthly tag cloud in Fig. 2 (left) shows the page-triggered keyword "smoking" as the most important topic. Analyzing frequently co-occuring keyphrases for "smoking" (see Fig. 2, right) yields (i) semantically related terms like "quit", "cigar", "cvd" (cardiovascular disease) and (ii) keyphrases indicating a personal connection (e.g., "husband", "father", "worries"), both mostly user-triggered. Apparently, many users are focusing on personal, rather than professional, aspects of the domain. As an actionable conclusion, it might be of use to trigger more treatment-related topics in order to shift engagement from patients to professionals.

4 Conclusion and Future Work

One of the biggest benefits social media channels add to modern marketing is the direct feedback on the content strategy. In order to help refine such a strategy, we have proposed Text Insights, a tool to analyze social media content on a textual level. Apart from content evaluation and planning, we have demonstrated that the system may also enable page owners to gain an understanding of trends in their communities. Since the very nature of social media is public, this can also be applied to a competitors' presence in order to compare content strategies.

As future improvements, we plan to include spam or language detection, since some contributions have shown to skew part of the generated tag clouds. The linguistic analysis might benefit from clustering synonymous keyphrases [1].

References

1. D. M. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
2. R. Miller and N. Lammas. Social media and its implications for viral marketing. *Asia Pacific Public Relations Journal*, 11(1):1–9, 2010.
3. O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–390, Atlanta, Georgia, 2013.
4. G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.