VYSOKÁ ŠKOLA EKONOMIE A MANAGEMENTU
Nárožní 2600/9a, 158 00 Praha 5

SEMINÁRNÍ PRÁCE

KOMUNIKACE A LIDSKÉ ZDROJE

VYSOKÁ ŠKOLA EKONOMIE A MANAGEMENTU

Nárožní 2600/9a, 158 00 Praha 5

TÉMA A NÁZEV SEMINÁRNÍ PRÁCE

Téma SP: Marketingová komunikace

Název SP: Identification of the most and the least engaging topics a Facebook page talks about.

STUDIJNÍ MODUL / SEMINÁRNÍ BLOK

Marketing / Seminární blok 1

JMÉNO A PŘÍJMENÍ STUDENTA / STUDIJNÍ SKUPINA

Tadeáš Moravec / KLZ 26

KATEDRA

Katedra marketingu

POZNÁMKY A PŘIPOMÍNKY

PROHLÁŠENÍ STUDENTA

Prohlašuji tímto, že uvedená seminární práce je mnou vypracována samostatně, a že ke zpracování této seminární práce bylo použito pouze literárních pramenů v práci uvedených.

Datum a místo: 31. 5. 2017, Drahelčice

VYSOKÁ ŠKOLA EKONOMIE A MANAGEMENTU

Nárožní 2600/9a, 158 00 Praha 5

SOUHRN

1. **Cíl práce:**
   Identifikace nejvíce a nejméně sdílených témat, o kterých mluví facebooková stránka amerického časopisu Psychology Today.

2. **Výzkumné metody:**
   Identifikace témat je kvalitativní metoda obsahové analýzy. Pro tuto analýzu práce využívá algoritmů z oboru strojového učení („machine learning“). Seřazení témat je kvantitativní metoda.

3. **Výsledky výzkumu/práce:**
   V práci bylo identifikováno dvacet šest skupin facebookových příspěvků. U dvaceti čtyřech z nich je identifikováno, o jakém tématu pojednávají. u zbývajících dvou se identifikace nezdařila. Sociální média obsahují velké množství šumu a navzdory tomu, že byla přijata opatření omezující šum, některé výsledky nejsou úplně srozumitelné. Za nejsilnější metriku zapojení uživatelů je považován počet sdílení. Témata jsou seřazena od nejvíce sdílených (rodičovství) po nejméně sdílené (tištěný časopis). Všechna témata jsou představena v Tabulce 2.

4. **Závěry a doporučení:**
   Administrátoři facebookové stránky by měli směřovat omezené zdroje, dostupné pro tvorbu obsahu, k tématům, která generují více sdílení. Ačkoli počet komentářů není považováno za vhodnou metriku zapojení uživatelů, bylo by užitečné využít algoritmy zpracování textu, které práce představuje, pro porozumění nálady komentářů a posouzení, která témata generují pozitivnější a negativnější reakce. Představená matematická metoda může být využita i pro širší využití, než jenom identifikace témat na jedné facebookové stránce: získání náhledu užitečného pro návrh a vývoj produktů, analýza zákaznické skupiny konkurence či analýza širšího odvětví s několika různými facebookovými stránkami.

KLÍČOVÁ SLOVA

Práce s veřejností, Sociální média, Facebook, Machine learning

VYSOKÁ ŠKOLA EKONOMIE A MANAGEMENTU

Nárožní 2600/9a, 158 00 Praha 5

SUMMARY

1. **Main objective:**
   Identification of the most and the least engaging topics that the Facebook page of an American magazine Psychology Today talks about.

2. **Research methods:**
   Identification of the topics is a qualitative method of content analysis. The work utilizes Machine Learning algorithms for the analysis. Ranking the topics is a quantitative method.

3. **Result of research:**
   The work identifies twenty-six groups of Facebook posts. In twenty-four of them, the work identifies the topic the group of posts talks about, but it fails to identify the remaining two. Social media data is very noisy, and although specific steps were taken to reduce the noise, some results are not clear. The number how many times was a post shared on Facebook is considered the strongest engagement metric. The topics are ranked from the most shared one (parenthood) to the least shared one (the printed magazine). All topics are presented in Table 2.

4. **Conclusions and recommendation:**
   The Facebook page administrators should guide the limited resources they have available for creating the content to the topics that generate more engagement. Even though the number of comments that different topics received is not considered a good engagement metric, it would be useful to use the presented natural language processing tools to understand the mood of the comments and judge which topics generated more positive or more negative reactions. The presented mathematical approach can be utilized for other applications than just analyzing what topics a single Facebook page talks about: obtaining insights relevant to product design and development, analyzing a competitor‘s audience or analyzing a broader industry with several Facebook pages.

KEYWORDS

Public relations, Social media, Facebook, Machine learning

JEL CLASSIFICATION

C38, M31

Contents

List of Figures

List of Tables

1 Introduction

Social media are an increasingly important part of contemporary marketing, and many organizations maintain pages on Facebook, the largest social network today. Getting the audience engage and interact with the content being shared on Facebook is the key to using Facebook effectively. Like in other aspects of the organizations' operations, it is necessary to make sure that the resources spent on social media are used efficiently. In other words, organizations need to understand what content is more engaging and guide the resources to creating such content.

This paper's goal is to identify the most and the least engaging topics a Facebook page talks about. It presents a general approach that works on any Facebook page and shows the results on the Facebook page of an American magazine Psychology Today.

When a company starts its Facebook activity, it is relatively straightforward to understand what topics and what kind of content (text, video, etc.) do the users like the most. As the page grows in size, it gets more difficult. The most popular Facebook pages have tens of thousands of posts. Simply reading the whole page would be very tedious, and the analysis would probably not be great - it is simply too much information for a human to process. The author of this paper works as a software engineer, which allows him to use machine learning, a method that is somewhat uncommon in social sciences (Kosinski et al., 2016, p. 493). The approach presented in this paper is useful to administrators of larger Facebook pages.

Identification of distinct topics can be seen as a qualitative method of content analysis, even though it uses mathematical concepts. The method turns the whole Facebook page into one huge matrix, with rows representing distinct posts on the page. It then searches for similarities between the rows with a method called cluster analysis. Social media are a very noisy source of data, so there are several steps that attempt to reduce the noise as much as possible while still keeping relevant information. Ranking the topics that were found with the cluster analysis is a quantitative method.

The Theoretical Part of this paper shows how is this method relevant in the marketing field. It then introduces several concepts related to natural language processing and machine learning and joins them in the Method section to create a single process that can be used on any Facebook page. The Analytical part applies the Method to a particular page and presents the topics that were found.

The author is interested in human psychology. The proposed approach works on any Facebook page, even though some pages lend themselves to text analysis better than others. The paper shows the process on the Facebook page of Psychology Today, an American magazine, which aims to make psychology more accessible to the general public. The author is in no way affiliated with Psychology Today; it is just a field he is interested in. Hence, this work attempts to bridge software engineering, marketing communication, and psychology.

2 Theoretical and Methodological Part

This part first connects the topic with the broader marketing. It then describes and explains several concepts related to social media and analyzing the data that can be obtained from them. First, it explains what Facebook is and why is mining data from there useful. Then it proceeds to introduce several concepts related to machine learning and natural language processing. It concludes with a Method section that ties the concepts together in a single process that identifies topics a Facebook page talks about and that applies elementary statistics to learn which topics generate the most and least engagement.

2.1 Facebook

Human interactions and activities are increasingly conveyed by digital services and products connected over computer networks (Lambiotte, Kosinski, 2014, p. 1). A significant portion of these services is represented by social media; The Nielsen Company (2010) reports that Internet users spend more time on social media sites than on any other kind of services. Kaplan, Haenlein (2010, p. 60) define Social media as Internet applications, building on foundations of Web 2.0 and allowing its users to create and share content. As of December 2016, the largest social media site was Facebook, with 1.86 billion monthly active users (Company Info | Facebook Newsroom, 2017).

Facebook is characterized by two core features. First, as Gjoka et al. (2010, p. 2) describe, Facebook users create a network of “friends”, an undirected graph, where users represent the vertices and their “friendships” represent the edges. Second, as described by Sangvi (2006), the News Feed, a page that displays posts with information like profile changes, upcoming events or birthdays of the users’ friends. Every user has his or her own News Feed that is unique to him or her. Users can “share” various kinds of content, be it text, images, videos or Web links. These “posts” appear on their own profile pages, and also appear it their friends’ News Feeds.

In addition to profile pages, users can create similar pages dedicated to a certain cause, often to a company, but not limited to that (Champoux et al., 2012, p. 1). Activity shared by these pages shows in the users’ personal News Feeds, in the same way as the activity shared by their friends does. A key concept here is “following”: a user follows certain pages and users and sees their posts in his or her News Feed (Peters, 2011) - friends are followed by default. Users can express a positive reaction to a post with the "Like" button; they can comment on a post or “share” it (put it on their profile).

2.2 Social Media Marketing

Felix et al. (2017, p. 119) define social media marketing as the use of social media platforms and websites to promote a product or a service.

Marketing communication is presented as one of the Lauterborn’s (1990, p. 26) “four Cs” of marketing: consumer, cost, communication, convenience. The “four Cs”, as explained in the Lauterborn’s article, roughly correspond to the “four Ps” presented by other authors (Kotler, Armstrong, 2015, p. 12): product, price, promotion, place. In this sense, marketing communication can be understood in a similar sense as promotion and Kotler, Armstrong (2015, p. 408) see it in the same way; they mention that promotion mix is also called marketing communication mix.

There are five primary promotion tools defined by Kotler, Armstrong (2015, p. 408):

1. Advertising

2. Sales promotion
3. Personal selling
4. Public relations
5. Direct marketing

They explain that public relations is a mass promotion tool that can be used to promote such diverse things as products, people, places, or even nations.

They describe several public relations tools (Kotler, Armstrong, 2015, p. 456):

1. News
2. Speeches
3. Special events like conferences, press tours, grand openings, hot air balloons, etc.
4. Written materials
5. Audiovisual materials
6. Corporate identity materials like brochures, business cards, etc.
7. Public service activities
8. Web: web sites, blogs, social networks like YouTube of Facebook

They understand the importance of social networks like YouTube, Facebook, and Twitter. They claim that the essentials of public relations - storytelling and sparking conversation - corresponds well with the nature of social media. Even though other Web tools are important as PR vehicles, especially the company's website, this paper targets Facebook, the most popular social network in 2017 (Company Info | Facebook Newsroom, 2017).

Roughly 25% of all people in the world use Facebook (Company Info | Facebook Newsroom, 2017) and social media play a central role in many individuals' lives (The Nielsen Company, 2010). As a result, social media provide a highly potent tool for marketing. Bowden (2014) reports, that social media marketing has brought retailers 133% increase in revenues in 2014, and in the same year, over 80% of business executives identified social media as an integral part of their business (Bennett, 2014).

Jansen et al. (2009, p. 2170) describe how social media users can "repost" comments made by others about a product or a service that is being promoted and they observe that this happens quite frequently on some social media sites. By repeating the message, the user's connections see the message, which then reaches more people. Thus, companies only put information about the product online once, and by getting repeated by the social media users, more traffic is brought to the product or business (Assaad, Gomez, 2011, p. 20).

Hanna et al. (2011, p. 272) claim that traditional media can be cost-prohibitive to many companies. They assert that in contrast, a well executed social media strategy does not need a huge budget.

Companies can analyze the customers' voices generated in social media. In this sense, social media are a relatively inexpensive source of marketing intelligence. The feedback can be used by marketers and managers to track and respond to problems identified by the customers. Constantinides et al. (2008, p. 17) describe, how it can be utilized to detect new market opportunities.

Mahapatra (2013) observes that brands that are less active on social media tend to show up less on Google searches. The key to successful social media marketing is to get consumers and potential customers to engage online (Evans, 2011). Hence, generating content that social media users would find engaging, plays a crucial role in modern marketing. Understanding, which type of content users like and share the most, plays a pivotal role in that.

2.3 Social Media Mining

Kotler, Armstrong (2015) do not mention using social media as a source of data for marketing research. Daymon, Holloway (2011, p. 288) hint about it, but they do not elaborate either. Meanwhile, Zafarani et al. (2014, p. 10) and Kosinski et al. (2016, p. 1) emphasize that social media have a huge potential for social science research.

The unprecedented availability of digital footprints (not limited to social media), combined with the power of modern computers, offers great opportunities for social sciences (Lazer et al., 2009, p. 723). Zafarani et al. (2014, p. 16) define Social Media Mining as the process of representing, analyzing, and extracting actionable patterns from the social media data.

Henrich et al. (2010, p.111) show, how mining big data from social media allows us to discover patterns, which might not be visible in smaller samples. They describe how it helps to reduce sampling errors typical in social science studies. They emphasize another major problem in social science studies: the studies rely on small samples, composed of mostly female students, who are disproportionately “*WEIRD - Western, Educated, Industrialized, Rich, and Democratic*” (Henrich et al., 2010, p. 111). They conclude that social media like Facebook are used by a more diverse population, which helps to address this problem.

Social media data is, without a doubt, big. Yet, when we zoom at the individuals, we often have very little specific information. Zafarani et al. (2014, p. 17) call this phenomenon the Big Data Paradox. They describe that it is necessary to exploit the characteristics of social media - use its many dimensions, many sources to aggregate data with enough statistics for useful analysis.

They also emphasize, that by its nature, social media contain a significant portion of noisy data. They make two important observations. First, blindly removing noise can make the problem worse, as valuable information could be eliminated as well. Second, the definition of noise can be complicated and relative, because it depends on what problem is the researcher trying to solve. The notion of noise is a paramount finding that plays a crucial role in the Analytical part of this paper.

2.4 Natural Language Processing

Liddy (2001, p.1) describes natural language processing (NLP in short) as computer analysis of text that relies on a set of theories and technologies. She notes, though, that there is no single universally agreed-on definition. There are some aspects such definition should contain, however. Henceforth, she defines NLP as "*a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.*" She emphasizes the word processing, as opposed to understanding. Even though NLP was originally referred to as natural language understanding in the early days of artificial intelligence research, she concludes that it is well understood now that humankind has not accomplished true understanding yet.

More recently, Bird et al. (2009, p. ix) describe natural language processing in a wider sense. In their understanding, it covers all kinds of computer manipulation of natural language, from simply counting word frequencies to understanding complete human expressions, at least to be able to respond to them in a meaningful way.

They observe that NLP is becoming increasingly common. Phones use predictive text or handwriting recognition. Web search engines can provide answers found in an unstructured text. Machine translation helps us to understand texts written not only in a language we do not speak but also in a script we cannot read. Sentiment analysis is an important part of many online retail applications.

Dumais et al. (1998, p. 148) propose another application of NLP: document categorization, the assignment of documents into several categories based on similarity of their content. It is possible to use predefined categories as well as let the system figure them itself.

2.5 Bag of Words

One way to represent a set of documents digitally for statistical processing is the bag of words method (BoW in short). BoW is an orderless representation; we only keep frequencies of words from a dictionary (Salton, McGill, 1983, p. 208).
As an example, let us have two documents (sentences):

1. One way to represent a set of documents digitally for statistical processing is the bag of words method.
2. Bag of Words is an orderless representation; we only keep frequencies of words from a dictionary.

There are 25 unique words in these two sentences. For example, word “one” is once in sentence 1 and zero times in sentence 2. Word “frequencies” is zero times in sentence 1 and once in sentence 2. Word “of” is twice in each sentence. We can construct a table like this for each of the 24 unique words.
From a mathematical point of view, sentence is represented as a vector (first order tensor, actually) with as many dimensions as there are distinct words in the whole corpus. And the whole corpus is represented as a vector of these sentence-vectors (second order tensor). This vector of vectors can be expressed by a matrix called the term-document matrix, with a column for every word and a row for every document, as described by Xu et al. (2003, p. 269). Note that in that paper, they use a transposed matrix and have rows for words and columns for documents.
There are several problems with the bag of words approach, however:

1. Inflection gives us different words with almost the same meaning, e.g. "represent" and "representation".
2. Certain word combinations convey more information than the individual words. "Fast" and "food" without order mean something different than "fast food".
3. Very common words appear disproportionately important. In the example above, the word "of" is present twice in every sentence, but carries little useful information about the sentences’ meaning.
4. All information conveyed by the grammar gets lost. For instance whether was the sentence a question or if an adverb was placed at the beginning of the sentence to be emphasized.
5. BoW treats synonyms as different words. Polysemy is a problem too.

Lovins (1968, p. 22) proposes to address the first problem by replacing all words by their stems. As an example, even though the sentences talk about the same thing, words "represent" and "representation" make them appear different in BoW. When reduced to the stem, "represent", the model correctly shows that both sentences contain a single instance of the same concept.
To address the next problem, Broder et al. (1997, p. 1161) propose the use of n-grams (called “shingles” in the original paper), sequences of n items from a given document. In addition to single words "one", "way", "to", "represent", we use bigrams "one way", "way to", "to represent". It is possible to go further and use trigrams "one way to", "way to represent" and possibly even more-grams. This method makes it possible to capture meaning in specific common collocations, but it does not help in understanding the grammar. While most n-grams do not contain

any useful information and most of them is rare, some are important. Using n-grams dramatically increases the amount of both data and noise. In turn, it is possible to reduce the noise by keeping only those words and n-grams, which are present in multiple documents in the corpus. To work around the disproportionate importance of very common words, a database of these words can be developed, and they can be simply skipped. Such database is called stopwords and usually contains words like articles or prepositions (Leskovec et al., 2015, p. 75). Common words, capturing little information, however, are often specific to the corpus. Essentially, there are different common words in civil engineering and marketing.

2.5.1 Tf-idf

Sparck Jones (1972, p. 13) notes that the peculiarity to of a term can be expressed as the inverse value of the number of documents that contain the term - the fewer documents mention the term, the more specific it is. Plain bag of words method assigns term frequency to each word. Multiplying the term frequency (tf), like in plain BoW, by the inverse document frequency (idf), i.e. the inverse number of documents that contain at least one occurence of the word, results in a metric called tf-idf. This metric gives more weight to terms that are peculiar to few documents and less weight to common terms (Leskovec et al., 2015, p. 8).

Beel et al. (2016, p. 330) claim that tf-idf is the most popular weighting scheme in contemporary research and applications, and that 70% of recommender systems in digital libraries utilize tf-idf. Stemming, n-grams, stop words and tf-idf together provide a powerful tool for identification of relevant words and phrases. However, it does not help to understand the grammar, nor does it address the synonymy and polysemy problem. Liddy (2001, p. 4) emphasizes that grammar does not lend itself easily to computational implementation, but there are statistical tools that help in understanding synonyms and polysemes.

2.6 Latent Semantic Analysis

Dumais (2004, p. 191) describes latent semantic analysis (LSA) as the use of statistical computation to extract the meaning of words and their contexts. LSA serves three purposes. First, it helps to disambiguate synonymy and polysemy. Second, it can reduce noise in the documents that are being analyzed. Third, it significantly speeds up later processing.

To understand the mathematics behind LSA, it is necessary to know what is singular value decomposition (SVD). SVD is a method to find three matrices, whose dot product is an approximation of the original matrix (Banerjee, Roy, 2014, p. 371). In the matrices generated by SVD, rows and columns are ordered from the most to the least significant, and the user can decide how many rows and columns to use, hence how close will the approximation be. Latent semantic analysis is the use of singular value decomposition on the term-document matrix obtained from BoW or tf-idf Dumais (2004, p. 192).

The inevitable question of how many dimensions should the system use is answered by Bradford (2008, p. 160). He claims that the number is limited by the number of documents that are being processed. As a general rule, hundreds of thousands of documents will work best best with around 300 dimensions; millions of documents will result in best effects with around 400 dimensions. Smaller number of dimensions is more useful for broader comparison of concepts in the text, whereas more dimensions will enable more relevant (and specific) concepts.

A seventeen years old student at the end of the high school knows about 80,000 words (Hirsch, 2003, p. 16). It can be expected that the term-document matrix can have up to this number of columns and n-times more if it contains n-grams. Latent semantic analysis reduces it to several hundred by mapping similar concepts to a single dimension.

Coincidentally, because LSA can reduce noise, it can work around the need for stemming (see above), but this does not mean that stemming should not be performed. Bypassing the need for stemming is especially useful when a researcher works with a language that does not have any ready-made tools for it.

2.7 Cluster Analysis

Bailey (1994, p. 34) describes cluster analysis as an effort to group objects into classes on the basis of their likeness on one or more traits. Estivill-Castro (2002, p. 65) observes that the notion of a cluster is not precisely defined and depends on specific use. He explains that for this reason, there are many methods how to find them.

The term-document matrix, obtained from the bag of words or tf-idf (and possibly reduced by the LSA) can be understood as a set of vectors, where each vector represents a point in a highly dimensional space. Asserting that similar items are close to each other, cluster analysis can be formulated as an attempt to assign labels to individual items in a manner, that minimizes distances between items with a single label and maximizes distances between items with different labels at the same time. This leads to using "hierarchical clustering" algorithm, which tries to minimize distances between data points (Rokach, Maimon, 2010, p. 278). Hierarchical clustering requires its users to specify the desired number of clusters upfront.

There are several methods to evaluate the cluster distribution quality. One of them is the silhouette coefficient, described by Rousseeuw (1987, p. 53). Silhouette coefficient is a measure how similar are objects within one cluster and how well are different clusters separated. It ranges from -1 (very poor result) to 1 (very good result). Values around zero suggest overlapping clusters.

2.8 TextRank

TextRank, an algorithm introduced by Mihalcea, Tarau (2004, p. 1) is an approach to text summarization. Mani (2001, p. 23) defines text summarization as an attempt "*to take an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user's or application's needs.*" He describes two core methods for text summarization: extraction - trying to identify the most specific words or sentences; and abstraction - trying to create a summary that does not contain text present in the text being summarized. TextRank is an extraction method.

It works analogously to PageRank, the algorithm that underpins the Google search engine (Brin, Page, 1998, p. 2). Page Rank (Brin, Page, 1998, p. 2) is a graph algorithm, operating on a graph of web pages (the vertices) and web links between them (the edges). It assigns relative importance to a web page by measuring how many web links point to it. It is not a simple count of the web links, however; it takes the importance of the source pages into account as well.

As described by Mihalcea, Tarau (2004, p. 1), sentences in a document can represent vertices in a graph and their similarity can represent the (weighted) edges. This is analogous to the graph of web pages and web link counts. Applying PageRank on such graph is called TextRank, and it ranks sentences by their relative "importance" in the text.

One way to measure the similarity between sentences is by measuring the distance of the respective tf-idf vectors, possibly with the LSA applied as well, in the same manner as in the cluster analysis (see above). In the paper where they introduce TextRank, Mihalcea, Tarau (2004, p. 7) propose a new metric based on the number of common words between two sentences. They notice that simple count of common words promotes long sentences. Their new metric normalizes the value by dividing the number of common words with a sum of logarithms of lengths of the two sentences.

Taking several of the most “important” sentences can yield a summary of the text. The sentences can be taken from arbitrary points of the original text, however, so they do not follow each other. It makes sense to present the summary in points, rather than in a block of text.

2.9 Method

The literature research draws from two major areas: marketing publications, particularly Kotler, Armstrong (2015), and machine learning publications, particularly Zafarani et al. (2014). It presents several concepts related to natural language processing and machine learning: tf-idf, latent semantic analysis, cluster analysis and TextRank. This section joins them in a single workflow.

The method itself has three major parts. First, get the data from Facebook. Second, process it with machine learning algorithms described above. Third, analyze the results.

Downloading data directly from the Facebook Web application violates the Terms of Service (Terms of Service, 2017). Facebook provides a special interface for this purpose, called the Graph API (API stands for application programming interface) (Graph API, 2017). This tool provides access to a wide range of Facebook functionality while protecting the users’ privacy with fine-grained access control policies. This paper does not describe the details of getting the data from there, because the Graph API (2017) documentation is very clean and easy to understand and the methods occasionally change, which would render this section out-of-date regardless.

As the Theoretical part structure suggests, there are four steps in processing the data:

1. Turn the Facebook page into a matrix using the tf-idf method.
2. Reduce the matrix size using the latent semantic analysis.
3. Perform a cluster analysis on the result.
4. Use the TextRank algorithm to extract the key sentences from each cluster found at the previous step.

The following paragraphs elaborate on the steps in more detail.

First, it is necessary to preprocess the raw data a little. Remove punctuation, replace newline characters with spaces, remove stopwords, replace all words with their stems and strip web links. The tf-idf method then turns the whole Facebook page to a matrix, with rows corresponding to individual posts and columns representing words. Tf-idf can create additional columns for the n-grams of words. This paper uses n-grams for n from 1 to 5. As described in the LSA chapter above, there are potentially tens of thousands of distinct words, and with n-grams, there can be even hundreds of thousands of columns. To reduce this amount, this paper only uses those words (and n-grams) that appear in at least three posts, as suggested by Joachims (1996, p. 2). He also notes that this removes most of the spelling errors.

The second step reduces the number of columns in the matrix, using the latent semantic analysis. This step turns the data into a much smaller matrix, where rows still correspond to individual messages, but there are much fewer columns. It is necessary to specify the target number of dimensions, however. Bradford (2008, p. 160) suggests three hundred if there are hundreds of thousands of posts and four hundred if there are millions of posts. There are probably not that many posts on a Facebook page, so one hundred looks reasonable.

The rows can be regarded as points in a one-hundred-dimensional space. In step three, the hierarchical clustering algorithm searches for points that are close to each other, measured with the regular Euclidean distance (i.e. computed by the Pythagorean theorem). The cluster analysis gives us a list of cluster assignments, where every row from the matrix is assigned a label - the number of the cluster it belongs to.

The hierarchical clustering algorithm requires the user to specify the number of clusters to search for. It is somewhat arbitrary, but from the nature of the task, it can be assumed that the number should lie between 15 and 30. On one extreme, a Facebook page probably talks about no less than 15 topics, and if it does not, it makes sense to split some of them. On the other extreme, analyzing more than 30 topics would be rather tedious, beating the purpose of the automatic tool. A reasonable approach is to perform the cluster analysis for all numbers between 15 and 30 and select the distribution with largest silhouette coefficient.

The fourth step, extraction of the key sentences from each cluster, is necessary because what the clustering algorithm finds similar might not always be obvious to a human reader. After all, as Zafarani et al. (2014, p. 17) emphasize, social media contain a big portion of noisy data. Here is where the TextRank algorithm comes into play. The graph, on which TextRank operates, could be constructed from either the whole Facebook posts, or the posts can be split into sentences, and the algorithm can operate on the sentences. Using the whole posts would result in longer points in the summary, but they could be potentially less idiosyncratic, because a single sentence, being shorter, is more focused than a post that can contain several sentences. This paper splits them into sentences first. Selecting five of the most characteristic sentences should be enough to get a grasp of what topic do they have in common - Mihalcea, Tarau (2004, p. 7) use 100 words long summaries, which roughly corresponds to five sentences. Besides the sentences, i.e. the vertices in the graph TextRank operates on, TextRank requires a similarity metric which represents weights of the edges in the graph. This paper uses the metric proposed by Mihalcea, Tarau (2004, p. 7), see the TextRank chapter above for details.

With the posts assigned to clusters, computing average numbers of likes, comments, and shares is trivial and so is comparing it to page-wide statistics. The last step is to order the topics by all three metrics and identify what topics are the most and the least engaging, where engagement can mean the number of likes, comments as well as shares.

It is important to check when was a bulk of posts in the cluster written. If left unchecked, the marketer could be lead to believe that a topic is very popular, even though it was only interesting several years ago. As a simple way to get this information, it is possible to turn dates into numbers by computing a difference between the time when was the message published and some fixed point in time. Then compute the standard deviation of these numbers and subtract and add it to the mean. This results in two dates between which most of the posts was written.

This paper concludes with a summary of the most and the least engaging topics that were identified, with a discussion of some limitations of the Method, and with hints of other use for this technology.

3 Analytical Part

The analytical part goes through the Method section from the previous chapter and presents the results. The Method has three major parts: get the data from Facebook; process it with machine learning; and analyze the results. This chapter is split a bit differently, however. There is less emphasis on the first two steps because they are already covered in detail in the previous section. Presenting what topics the machine learning algorithms found takes most of the space. The Analytical Part continues with a summary of the topics and their order in terms of different engagement metrics. And it wraps up with a small discussion of the least engaging topics and with hints of directions for future research.

3.1 Data from Facebook

As of 5/23/2017, the current version of the Facebook Graph API is 2.9 (Graph API, 2017). According to its documentation, the correct way to get the whole Facebook page is this query:

```
https://graph.facebook.com/v2.9/psychologytoday?fields=
posts{message,link,likes.limit(0).summary(true),
comments.limit(0).summary(true),shares,description,
name,created_time}&access_token=<access␣token>
```

This query returns posts the page shared in JSON format, which is suitable for subsequent machine processing. Users need to supply their own access token when running this query in the <access token> field, however.

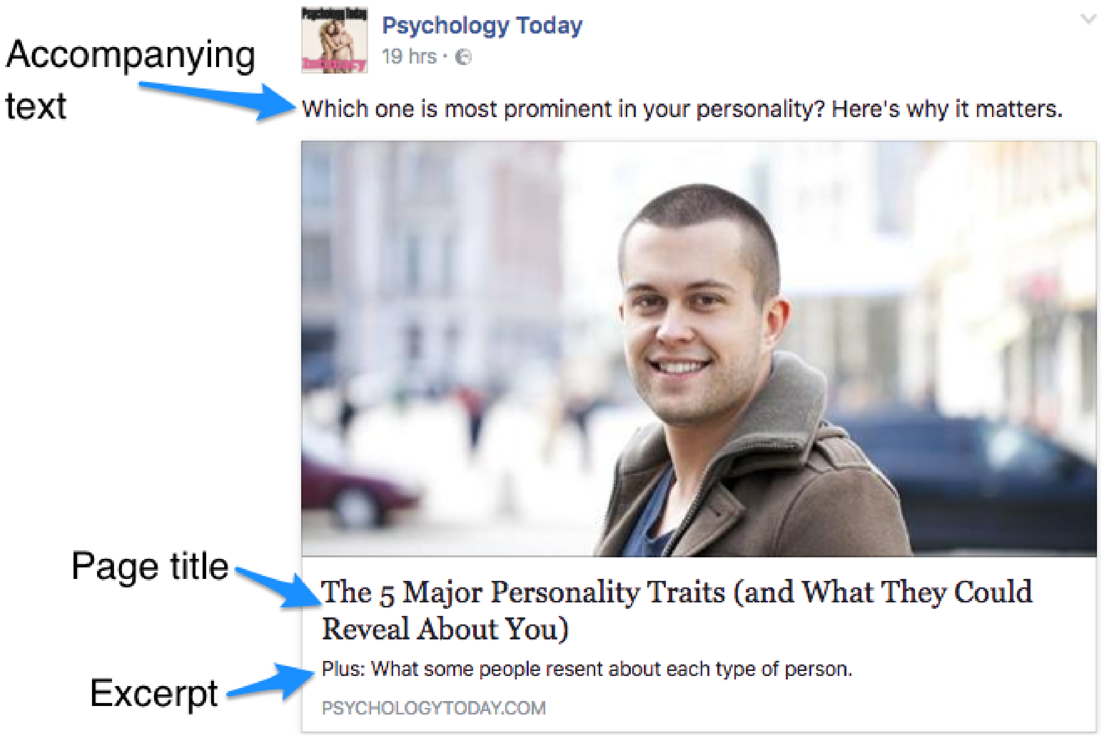
When a user shares a web link to Facebook, in addition to the accompanying text, Facebook displays a preview of the linked web page. This preview usually contains the title of the page, and a very short excerpt from the linked page text, together with a picture taken from the page. In these cases, this paper concatenates the accompanying text with the page title and the excerpt, because this is what the users immediately see when they stumble on the post. In Figure 1, this leads to message "Which one is the most prominent in your personality? Here's why it matters. The 5 Major Personality Traits (and What They Could Reveal About You) Plus: What some people resent about each type of person."

As of 5/23/2017, Psychology Today Facebook page shared 7658 posts. Only those newer than 1/1/2012 are considered though, for two reasons. First, recommendations for marketing communication should be based on recent trends. Second, engagement statistics for the posts are computed. The number of Facebook users increases over time, so older posts generated less engagement because there were fewer users, not because they were less popular. In the USA, the number of Facebook users reached a plateau around 2012 (Company Info | Facebook Newsroom, 2017), so 1/1/2012 is a good date to cut it off. There are 6237 messages newer than 1/1/2012, 4402 of which contain any text.

3.2 Processing the Data

Before vectorization, it is necessary to perform several sanitization steps on the data. After removing punctuation, replacing newline characters with spaces, replacing all words with their stems and striping web links, the sample post from Figure 1 looks like this: "which one is the most promin in your person here whi it matter the major person trait and what they could reveal about you plus what some peopl resent about each type of person"

Figure 1 An example of a Facebook post that links to a web page.



Psychology Today

19 hrs ·

Which one is most prominent in your personality? Here's why it matters.

The 5 Major Personality Traits (and What They Could Reveal About You)

Plus: What some people resent about each type of person.

PSYCHOLOGYTODAY.COM

Source: Psychology Today - Home (2017), with the author's annotations.

After sanitizing all posts in this manner, they are transformed into a matrix with the tf-idf method. N-grams are used as well, for n from one (individual words) to five. Tf-idf generates a matrix with 4402 rows and 219 705 columns. This is reduced significantly by only considering terms (including n-grams) that are present in at least three posts. This reduces the number of the columns to 5888. Each row is mostly zeros, with several non-zero rational numbers between them. The latent semantic analysis further reduces this matrix to 100 columns, keeping 4402 rows.

Cluster analysis is performed on this matrix, with the number of clusters for all numbers between 15 and 30 (inclusive) and silhouette score is computed for each distribution. It turns out that the best number of clusters is 26 with silhouette score 0.020 (all results in Table 1).

3.3 Page Statistics

Psychology Today Facebook page has 7 461 278 fans and shared 4402 posts that can be analyzed. The remaining posts do not contain any text (e.g. video or photo only), or they are older than 1/1/2012.

The posts have, on average, 2278 likes, 110 comments and they were shared 1018 times.

3.4 Topics Found

Topic 1

It looks like the underlying theme is *parenthood*.

70 posts

Summary:

Table 1 Silhouette scores for different numbers of cluster.

Cluster	Score
15	0.0164
16	0.0165
17	0.0168
18	0.0168
19	0.0167
20	0.0177
21	0.0180
22	0.0176
23	0.0171
24	0.0182
25	0.0197
26	0.0200
27	0.0184
28	0.0184
29	0.0187
30	0.0197

Source: The author's research.

- Why Parents Really Get Angry at Their Kids (and Why It's Pointless) Kids do exactly what they're meant to do, and expert says.
- Not Naughty: 10 Ways Kids Appear to Be Acting Bad But Aren't ... and why sometimes it's the parents who are really responsible.
- 3 Reasons Why Parents Let Their Kids Bully Them.
- 'Good enough' parents may produce more resilient kids than 'super' parents, according to new research.
- What We Really Needed From Our Parents (and What Our Kids Need From Us) Many of our fondest childhood memories are of getting in trouble.

Average likes: 2418
Average comments: 182
Average shares: 1822
Roughly from 2014/5 to 2017/2

Topic 2

It looks like the underlying theme is *problematic people*.
59 posts
Summary:

- 8 Ways to Deal With the Most Toxic People in Your Life.
- The Best Way to Deal With Controlling People.
- 4 Ways to Deal With Boring People.
- Yes, there are ways to deal calmly with the people who infuriate you.
- 5 Ways to Deal with Angry People.

Average likes: 2734

Average comments: 135
Average shares: 1540
Roughly from 2014/9 to 2016/11

Topic 3

It looks like the underlying theme is *mental strength*.
58 posts
Summary:

- 7 Ways Mentally Strong People Bounce Back.
- 5 Ways Mentally Strong People Bounce Back From Rejection.
- 8 Ways That Mentally Strong People Gain Financial Freedom.
- 7 Ways Mentally Strong People Deal With Stress.
- 5 Ways Mentally Strong People Conquer Self-Doubt.

Average likes: 2884
Average comments: 109
Average shares: 1487
Roughly from 2014/8 to 2017/2

Topic 4

It looks like the underlying theme is *bad things people say*.
92 posts
Summary:

- 5 Reasons Why Some People Will Never Say "I'm Sorry" 4.
- Why do well-meaning people say insensitive things?
- Why do otherwise well-meaning people say insensitive things?
- The 3 Most Common Excuses for Cheating (and Why They're Bogus) ... and why an expert says that every couple needs to draw the line on what counts as infidelity.
- The 3 Things You Should Never Say to Your Partner.

Average likes: 2795
Average comments: 127
Average shares: 1240
Roughly from 2014/10 to 2016/9

Topic 5

It looks like the underlying theme is *when things go wrong*.
108 posts
Summary:

- 5 Ways Relationships Can Go Wrong (and 3 Ways to Fix Them) When things start to go wrong, can you avoid a negative loop?
- When things go wrong, do you take the blame, or blame anyone you can find?
- 3 Thoughts That Can Hold You Back (and How to Let Them Go) 3.
- When things go wrong, you might find a clearer path to success.

- They make themselves responsible when things go wrong.

Average likes: 2776
Average comments: 98
Average shares: 1219
Roughly from 2014/2 to 2016/6

Topic 6

It looks like the underlying theme is *happiness*. Most of the posts were written at the beginning of 2014.
73 posts
Summary:

- 5 Things Happy People Do Every Day (and You Can, Too) Money can't buy happiness, unless you spend it the right way.
- What do happy people do that makes them so happy all the time?
- Why Trying to Make Everyone Happy Can Make You Miserable.
- Why Seeking Happiness Every Day Could Make You Miserable.
- Is Making Other People Happy Making You Miserable?

Average likes: 2735
Average comments: 106
Average shares: 1215
Roughly from 2014/1 to 2014/1

Topic 7

It looks like the underlying theme is *lies*.
39 posts
Summary:

- 7 Ways to Really Tell if Someone Is Lying to You.
- Think you can tell when someone's lying?
- Psychological research has made one thing pretty clear: human beings are much better liars than lie detectors.
- How the Experts Spot Liars (and How You Can, Too) Our mind and body literally betray us when we lie.
- This may be the most reliable way to detect someone's lies.

Average likes: 2766
Average comments: 118
Average shares: 1185
Roughly from 2014/4 to 2016/6

Topic 8

It looks like the underlying theme is *how to stop worrying*.
99 posts
Summary:

- 6 Ways to Stop Worrying About the Things You Can't Control.
- 2 Ways to Stop Worrying and Overcome Anxiety.
- Why can't you stop thinking about things you don't want to think about?
- 4 Ways to Stop Worrying.
- Do you know anyone who can't stop worrying about what other people think of them?

Average likes: 2309
Average comments: 85
Average shares: 1185
Roughly from 2014/2 to 2016/9

Topic 9

It looks like the underlying theme is *narcissism*.
64 posts
Summary:

- New research explains why things go wrong when your partner is a narcissist.
- What Makes Narcissists Angry (and Why) New research into anger style and instincts for revenge.
- New research explores what it's really like to be in a committed relationship with a narcissist.
- New research on how different types of narcissists approach relationships.
- Why Narcissists Want to Make Their Partners Jealous.

Average likes: 2340
Average comments: 232
Average shares: 1087
Roughly from 2014/12 to 2017/2

Topic 10

It looks like the underlying theme is *new studies results*.
70 posts
Summary:

- A new study shows that you really do make your own breaks.
- If a man wants to meet women, a new study shows, he can't go wrong with this strategy.
- A new study shows why.
- A new study finds that analyzing Facebook likes reveals more than you might expect.
- A new study finds that our desirability may hinge on how many previous partners we've had.

Average likes: 2447
Average comments: 190
Average shares: 1065
Roughly from 2014/3 to 2016/1

Topic 11

This cluster is extremely large (1211 posts) and it looks like it contains a big portion of noise. It looks like the underlying theme could be relationships. But almost all posts on the whole page talk about relationships in one way or another (it is about psychology after all), so this is apparently mostly *noise*.
1211 posts
Summary:

- How You Really See Your Partner (And Why It Matters So Much) Research shows that looks matter, but not in the way most of us think.
- 5 Hidden Biases You Probably Have (Even If You Think You Don't) ... including why we favor more attractive people (and are sure our partner is one of them).
- 4 Reasons to Make New Year's Resolutions (and 3 Reasons Not to) ... and why some people just need to resolve to stop making excuses.
- 3 Reasons Why People Avoid Talking About "The Relationship" Sometimes we push to define things too soon.
- Why You Should Never Tell Someone "I Need You To..." In a relationship, the phrase can mean two very different things.

Average likes: 2295
Average comments: 110
Average shares: 1045
Roughly from 2014/4 to 2016/9

Topic 12

This cluster is extremely large (881 posts) and it looks lik it contains a big portion of noise as well. Like in Topic 11, it looks like this cluster is mostly *noise*.
881 posts
Summary:

- To make better relationship decisions in the year ahead, take to heart what many have learned the hard way.
- To make better relationship decisions in the future, take to heart what many have learned the hard way.
- 10 Ways to Break Out of a Rut (Right Now) Unplanned moments make us feel happier and more in control, research finds.
- 5 Ways to Make Sure You Keep Your Relationship Happy.
- 3 Surprising Ways Dogs Make Your Relationships Better.

Average likes: 2275
Average comments: 91
Average shares: 1042
Roughly from 2014/4 to 2016/8

Topic 13

It looks like the underlying theme is *friendship*.
97 posts

Summary:

- Need to give a friend some advice, but want to keep your friendship?
- The 9 Simple Ways to Keep Your Closest Friends Close (and Why It Matters So Much) Make real-world plans, be there during hard times, and always—always—return their calls.
- Can friends become lovers without risking their friendship?
- "A true friend is consistently willing to put your happiness before your friendship."
- Why Our Best Friends Can Be So Good at Hiding Things From Us.

Average likes: 2604
Average comments: 121
Average shares: 1029
Roughly from 2014/6 to 2016/6

Topic 14

It looks like the underlying theme is *love*.
121 posts
Summary:

- Why does criticism from a loved one hurt so much more than from anyone else?
- One reason relationships are so hard is that falling in love is so easy.
- Why is it that our loved ones can hurt us more easily than anyone else?
- Why do more of us seem to be falling out of love with the idea of getting married?
- Two people could meet, fall in love, and start a life together, and only their families and perhaps closest friends would know.

Average likes: 2564
Average comments: 120
Average shares: 978
Roughly from 2014/1 to 2016/9

Topic 15

It looks like the underlying theme is *giving*.
95 posts
Summary:

- How giving time to others makes you feel like you have more time to give.
- "Giving time to others makes you feel like you have more time to give."
- If big family gatherings make you feel like a prisoner, these tips could help set you free.
- Giving a small donation at checkout can make you feel like you've "done your part".
- Ever feel like total strangers seem to understand you better than those closest to you?

Average likes: 2235
Average comments: 86
Average shares: 969
Roughly from 2014/3 to 2016/9

Topic 16

It looks like the underlying theme is *sleep*.
74 posts
Summary:

- Why You Need Your Partner to Get a Good Night's Sleep (and Vice Versa) Research finds a strong link between what goes wrong at night and what goes wrong the next day.
- Want to sleep better tonight than you did last night?
- Why Your Relationship Depends on a Good Night's Sleep.
- A good night's sleep really does clear your head.
- In a nation chronically wanting for rest, the gravest danger is not merely the molecular impact of sleep deprivation—but the simple fact that we no... Craving a good night's sleep?

Average likes: 2127
Average comments: 105
Average shares: 963
Roughly from 2013/12 to 2016/6

Topic 17

It looks like the underlying theme is *interesting new research*.
82 posts
Summary:

- New research shows why taking full responsibility feels better.
- Fascinating new research shows why they might.
- Fascinating new research shows how many of us see the physical world in vastly different ways.
- When we're thinking about sex, fascinating new research shows, we're much more likely to reveal deeper personal details about ourselves than we otherwise would.
- If your gut tells you that a partner might become unfaithful, new research shows, it may be wise to trust it.

Average likes: 2320
Average comments: 109
Average shares: 955
Roughly from 2014/7 to 2016/9

Topic 18

It looks like the underlying theme is *life*.
219 posts
Summary:

- The evidence is in: Singles are as happy as everyone else, live as long, and find as much meaning in life.
- If you are like millions of others, when it comes to making serious personal changes—the kind that can do you considerable good—you put on the... One way to be happier that's easier than acquiring new things: Imagine life without the things you already have.

- How to Get More Attached to the Place Where You Live (and Why It Could Save Your Life) Feeling rooted “is perhaps the most important and least recognized need of the human soul.” Love isn’t really blind.
- The Only Way to Make Positive Change in Your Life.
- Why Your Life Story Matters (and the 3 People You Need to Tell It to Now) Who you think you are has a huge impact on who you’ll turn out to be.

Average likes: 2249
Average comments: 91
Average shares: 937
Roughly from 2013/12 to 2016/9

Topic 19

It looks like the underlying theme is *questions to ask*.
67 posts
Summary:

- How to Answer the Question You Get Asked Most Often (and Why It Matters) It may seem automatic but it’s part of the glue of your relationships.
- 5 Questions to Ask Yourself Before Letting Go of a Friend.
- Asking the right questions can help someone raised to feel unlovable quiet their inner critic.
- 3 Questions You Have to Ask Yourself Before Buying Anything.
- ... and 14 questions to ask yourself before you strike back.

Average likes: 1930
Average comments: 76
Average shares: 907
Roughly from 2014/10 to 2016/10

Topic 20

It looks like the underlying theme is *romantic relationships*.
334 posts
Summary:

- The Top 10 Ways Couples Annoy Each Other (and What to Do About Them) If you’re not careful, the little things your partner does to irritate you can torpedo your relationship.
- Have you ever been in a relationship where one partner wants more space, but the other doesn’t want to give it?
- There’s one key factor that makes porn use damage a relationship, and it’s not how much a partner views it.
- Couples that stay together through one partner’s transition must navigate a complicated set of emotions.
- Does where you met your partner affect the success of your relationship?

Average likes: 2136
Average comments: 106

Average shares: 904
Roughly from 2014/7 to 2016/10

Topic 21

It looks like the underlying theme is *dreams*.
40 posts
Summary:

- New research uncovers why we dream what we dream, and when we're likely to dream it.
- New research into why we really dream about the things we do.
- This May Be Why We Recall Some Dreams (Especially Bad Ones) Studies open new ground in the search for clues to how the dreaming brain works.
- New research into who remembers dreams, and how our dreams affect us.
- The Surprising (and Fun) New Theory of Why We Dream.

Average likes: 2264
Average comments: 98
Average shares: 860
Roughly from 2014/5 to 2016/7

Topic 22

It looks like the underlying theme is *research on popular topics*. It looks somewhat similar to Topics 17 and it's not apparent why did the clustering algorithm separate these two clusters. Topic 17 is maybe focused on curiosities in new research, whereas this topic is more focused on what research shows about "old", but popular questions.
189 posts
Summary:

- New research reveals why time spent on social networks may be making you less happy.
- New research reveals who keeps the most secrets in relationships, and why.
- New research reveals why it's so hard, and a path to making it easier.
- New research reveals why it's so difficult and how we could crack the case.
- New research reveals better ideas to maintain successful relationships.

Average likes: 1942
Average comments: 123
Average shares: 772
Roughly from 2014/10 to 2016/12

Topic 23

It looks like the underlying theme is *gender differences*.
88 posts
Summary:

- New research reveals men and women's true (and very different) motivations.
- New research on how men and women judge each other's sexual history.

- Research reveals how men and women deal with jealousy in different ways.
- Research finds that romance offers men and women very different benefits.
- New research reveals how inked men are really perceived, both by both women and other men.

Average likes: 2026
Average comments: 198
Average shares: 743
Roughly from 2014/4 to 2016/11

Topic 24

It looks like the underlying theme is *Social Media*.
41 posts
Summary:

- A surprising study finds that people use social media to cheat on partners in more ways than most of us thought.
- Have you or a friend ever had a breakup take place online, or spill over into social media?
- A new study finds that social media may know how you're feeling before you do.
- Men and women get jealous in very different ways, and social media doesn't help.
- Have you lost someone and used social media as a way to remember or grieve for them?

Average likes: 1776
Average comments: 114
Average shares: 679
Roughly from 2013/12 to 2016/6

Topic 25

It looks like the underlying theme is *sex*.
88 posts
Summary:

- 4 Common Myths About Sex and Desire (and the Surprising Truth) Think it's abnormal to desire a dominant partner (or to be one)?
- The Surprising Secret to a Better Sex Life (and Relationship) ... and the 3 questions every couple must ask each other.
- When one partner wants to have sex much more often than the other (or much less), is there any way for them both to be happy?
- The new research on casual sex may surprise you.
- Why So Many Couples Are Having Less Sex.

Average likes: 1510
Average comments: 134
Average shares: 511
Roughly from 2014/7 to 2017/1

Topic 26

It looks like the underlying theme is *the printed magazine*.
43 posts
Summary:

- Psychology Today Covers 2013 to ... Help choose the next cover of Psychology Today by taking this brief survey!
- Help choose the next cover of Psychology Today (and offer your own feedback) with this quick survey!
- Help decide the next cover of Psychology Today and share your thoughts: Take this quick survey!
- Help decide the next cover of Psychology Today and share your thoughts by taking this quick survey!
- Help us choose the future cover of Psychology Today by taking this quick survey: https://www.surveymonkey.com/r/DMX5STH.

Average likes: 777
Average comments: 27
Average shares: 114
Roughly from 2012/7 to 2016/1

3.5 The Most and Least Engaging Topics

Commenting on a Facebook post can express both positive and negative attitude. It is not possible to simply look at the number of comments to judge if this is the kind of content the fans are interested in. On the other hand, the number of likes and especially the number of shares clearly indicate what the fans identify with. Sharing a post is considered the strongest indicator of engagement. When a user presses the Like button on a post, the post will be displayed in the users friends' News Feeds. When a user shares a post, it will not only show in the users friends' News Feeds, but the post will stay on the users profile page as well. The number of shares is usually significantly smaller than the number of likes (2278 vs. 1018 on average), which further supports this conclusion.

Table 2 presents all three metrics nonetheless.

It is apparent that the page administrators should focus on topics that generate most engagement, like parenthood, dealing with problematic people, mental strength, bad things people say or when things go wrong. The other end of the spectrum is not that obvious. The fact, that a topic does not generate that much engagement does not necessarily mean that the page should skip these posts. For instance, surveys about the magazine cover probably have other use than marketing communication. It does not change the fact, though, that from marketing communication point of view, it is more efficient to guide the limited resources to creating content that is more engaging and gets more eyeballs on the page.

With the exception of Topic 6 (happiness), the posts within individual clusters written do not appear to be exceedingly old. All of the first five topics contain posts that are relatively recent, and can be used as a guide for writing new engaging content.

The page-wide average number of shares across all posts that were considered is 1018. From the 26 topics found, the first 13 have above average number of shares, which is expected. Posts talking about the most shared topic (parenthood) were shared 1822-times on average, which is 179% of the page-wide average. Posts talking about the least shared topics (the printed

Table 2 Topics sorted by different metrics

Order	By likes	By comments	By shares
1	Mental strength	Narcissism	Parenthood
2	Bad things people say	Gender differences	Problematic people
3	When things go wrong	New studies results	Mental strength
4	Lies	Parenthood	Bad things people say
5	Happiness	Problematic people	When things go wrong
6	Problematic people	Sex	Happiness
7	Friendship	Bad things people say	Lies
8	Love	Research on popular topics	How to stop worrying
9	New studies results	Friendship	Narcissism
10	Parenthood	Love	New studies results
11	Narcissism	Lies	(Noise)
12	Interesting new research	Social Media	(Noise)
13	How to stop worrying	(Noise)	Friendship
14	(Noise)	Interesting new research	Love
15	(Noise)	Mental strength	Giving
16	Dreams	Happiness	Sleep
17	Life	Romantic relationships	Interesting new research
18	Giving	Sleep	Life
19	Romantic relationships	Dreams	Questions to ask
20	Sleep	When things go wrong	Romantic relationships
21	Gender differences	(Noise)	Dreams
22	Research on popular topics	Life	Research on popular topics
23	Questions to ask	Giving	Gender differences
24	Social Media	How to stop worrying	Social Media
25	Sex	Questions to ask	Sex
26	The printed magazine	The printed magazine	The printed magazine

Source: The author's research.

magazine) were shared 114-times on average, which is 11% of the page-wide average.

There are two interesting and unintuitive results, however: surprisingly little engagement with topics about sex (50% of average shares), social media (67% of average shares), and gender differences (73% of average shares). We can only hypothesize if these topics are indeed less attractive or if there is something else going on. Maybe sharing a post that talks about the dangers of social media on social media could appear awkward. Also, we can hypothesize that people are shy to share content talking about sex. We cannot know for sure from this analysis alone, however. From the point of marketing communication, both topics generate less engagement nonetheless, whatever is the underlying psychological reason. It could be an interesting area for future research.

While the number of comments a topic receives on average is not considered a good indicator of positive engagement for the marketing purposes, it might suggest that the topic is seen as controversial. The most commented-on topics are Narcissism, Gender Differences and New studies results, none of which belong to the most shared topics. Analyzing comments on these topics might be an interesting area for future research as well, altough it is less relevant from the marketing point of view and possibly more relevant from the psychology point of view.

3.6 Source Code

The author wrote a simple software tool to do the mathematical part of the analysis - tf-idf, the latent semantic analysis, the cluster analysis, and the TextRank summarization. The source code for this software is publicly available at https://github.com/tadeas/thematic. It does not bring anything new or interesting from the perspective of mathematics or computer science, however.

4 Conclusion

Sharing engaging content on social media helps companies with one of the key goals of public relations - building relationship with the companies' public. Additionally, by sharing such content, the marketers can reach more people, which has a positive impact on the broader promotion, one of the four key concepts of marketing. From the marketing point of view, it makes sense for the company to use its resources to generate content talking about the more engaging topics at the expense of the less engaging ones.

This paper's goal is to identify the most and the least engaging topics that the Facebook page of an American magazine Psychology Today talks about. The paper claims that the strongest engagement metric on Facebook is the number of shares a post received. Hence, it identifies these five topics as most engaging: Parenthood, Problematic people, Mental strength, Bad things people say, and When things go wrong. On the other end of the spectrum it identifies these five topics that generate the least engagement: Posts talking about the printed magazine, Sex, Social Media, Gender differences, and Research on popular topics. The whole list of the 26 topics identified in this paper is summarized in Table 2.

Sex and Social media being among the least engaging topics is somewhat surprising. It could be an interesting area for future research, albeit it is a question that belongs to the psychology field and has little to do with marketing communication.

It is important to understand that social media are a very noisy source of research data. As a consequence, there are clusters, where it is not possible to understand what do the posts have in common (topics 11 and 12). These clusters should probably be split into smaller ones, but that would result in too many topics to comfortably work with.

There is the opposite problem present as well: there are distinct clusters, which appear to talk about the same topic, namely topics 10 (new studies results), 17 (interesting new research), and 22 (research on popular topics).

The Most and Least Engaging Topics chapter notes that a high number of comments is not a good indicator of engagement because comments can express both positive and negative attitude. With natural language processing tools at our disposal, there could probably be a way to detect the mood of a comment and try to judge if the post generates more positive or more negative feedback. Analyzing the comments is another interesting area for future research.

This approach, which only groups Facebook posts by topics, can have other use, besides suggesting what topics should a page focus on.

First, it provides insights beyond marketing. Seeing what is and what is not interesting to a company's users can have a significant impact on the products and services development. In this sense, this allows hearing the voices of the silent majority of users (Mustafaraj et al., 2011, p. 103).

Second, it can be used on a competitor's page. It can give insight into what the competition focuses on, what their users find interesting and what they value. It is then possible to contrast the results with our company's marketing efforts, and the differences might identify areas where our company can vary from the competition.

Third, with minor modifications, it could be used to analyze multiple pages in a single or similar industry. Because every Facebook page talks about slightly different things, this could provide robust insights into what the customers find interesting beyond a single company customer base.

References

ASSAAD, Waad; GOMEZ, Jorge Marx. Social network in marketing (social media marketing) opportunities and risks. In: *International Journal of Managing Public Sector Information and Communication Technologies*. 2011. vol. 2, no. 1, p. 13. ISSN 2230 7958

BAILEY, Kenneth D. *Typologies and taxonomies: an introduction to classification techniques*. 1st ed. Thousand Oaks : Sage Publications, 1994. ISBN 0803952597

BANERJEE, Sudipto; ROY, Anindya. *Linear algebra and matrix analysis for statistics*. 1st ed. Hoboken : CRC Press, 2014. ISBN 9781482248241

BEEL, Joeran, et al. Research-paper recommender systems: a literature survey. In: *International Journal on Digital Libraries*. 2016. vol. 17, no. 4, p. 305-338.

BELCH, George E.; BELCH, Michael A. *Advertising and promotion: An integrated marketing communications perspective.* 5th ed. New York, NY : McGraw-Hill Education, 2003. ISBN 9781259548147

BENNETT, Shea. *Social Media Business Statistics, Facts, Figures & Trends 2014* [online]. April 25, 2014 [Accessed March 17, 2017] . Available from: <http://www.adweek.com/digital/social-business-trends-2014/>

BIRD, Steven; KLEIN, Ewan; LOPER, Edward. *Natural language processing with Python: analyzing text with the natural language toolkit.* 1st ed. Beijing : O'Reilly, 2009. ISBN 9780596516499

BOWDEN, Jason. *The Impact of Social Media Marketing Trends on Digital Marketing* [online]. March 17, 2014 [Accessed March 17, 2017]. Available from: <http://www.socialmediatoday.com/content/impact-social-media-marketing-trends-digital-marketing>

BRADFORD, Roger B. An empirical study of required dimensionality for large-scale latent semantic indexing applications. In: *Proceedings of the 17th ACM conference on Information and knowledge management*. 2008. p. 153-162. ISBN 9781595939913

BRIN, Sergey; PAGE, Lawrence. The anatomy of a large-scale hypertextual Web search engine. In: *Computer Networks and ISDN Systems*. 1998. vol. 30, no. 1. ISSN 0169-7552

BRODER, Andrei Z., et al. Syntactic clustering of the web. In: *Computer Networks and ISDN Systems*. 1997. vol. 29, no. 8, p. 1157-1166. ISSN 0169-7552

CHAMPOUX, Valerie; DURGEE, Julia; MCGLYNN, Lauren. Corporate Facebook pages: when "fans" attack. In: *Journal of Business Strategy*, 2012, vol. 33, no. 2, p. 22-30. ISSN 0275-6668

Company Info | Facebook Newsroom [online]. March 17, 2017. [Accessed March 17, 2017]. Available from: <http://newsroom.fb.com/company-info/>

CONSTANTINIDES, Efthymios; ROMERO, Carlota Lorenzo; BORIA, Miguel A. Gómez. Social media: a new frontier for retailers?. In: *European Retail Research.* Wiesbaden : Gabler Verlag / GWV Fachverlage GmbH, 2008. p. 1-28. ISBN 9783834980991

DAYMON, Christine; HOLLOWAY, Immy. *Qualitative research methods in public relations and marketing communications.* 2nd ed. London : Routledge, 2011. ISBN 9780415471176

DUMAIS, Susan T. Latent semantic analysis. In: *Annual review of information science and technology*. Medford, N.J. : Information Today, Inc., 2004. vol. 38, no. 1, p. 188-230.

DUMAIS, Susan, et al. Inductive learning algorithms and representations for text categorization. In: *Proceedings of the seventh international conference on Information and knowledge management*. New York, NY : ACM, 1998. p. 148-155. ISBN 1581130619

ESTIVILL-CASTRO, Vladimir. Why so many clustering algorithms: a position paper. In: *ACM SIGKDD explorations newsletter*, 2002. vol. 4, no. 1, p. 65-75. ISSN 1931-0153

EVANS, Dave. *Social media marketing: the next generation of business engagement.* 1st ed. Hoboken, N.J. : Wiley Pub. Inc, 2011. ISBN 9780470634035

FELIX, Reto; RAUSCHNABEL, Philipp A.; HINSCH, Chris. Elements of strategic social media marketing: A holistic framework. In: *Journal of Business Research*, 2017. vol. 70, no. 1, p. 118-126. ISSN 0148-2963

GJOKA, Minas, et al. Walking in Facebook: A case study of unbiased sampling of OSNs. In: *Proceedings - IEEE INFOCOM, (2010 06 15)*, 2010. p. 1-9. ISSN 0743-166X

Graph API [online]. March 18, 2017. [Accessed March 18, 2017]. Available from: <https://developers.facebook.com/docs/graph-api>

HANNA, Richard; ROHM, Andrew; CRITTENDEN, Victoria L. We're all connected: The power of the social media ecosystem. In: *Business horizons*, 2011. vol. 54, no. 3, p. 265-273. ISSN 0007-6813

HENRICH, Joseph; HEINE, Steven J.; NORENZAYAN, Ara. Beyond WEIRD: Towards a broad-based behavioral science. In: *Behavioral and Brain Sciences*, 2010. vol. 33, no. 2-3, p. 111-135. ISSN 0140-525X

HIRSCH, Eric Donald. Reading comprehension requires knowledge - of words and the world. In: *American Educator*. 2003. vol. 27, no. 1, p. 10-13.

JANSEN, Bernard J., et al. Twitter power: Tweets as electronic word of mouth. In: *Journal of the American society for information science and technology*. 2009. vol. 60, no. 11, p. 2169-2188. ISSN 1532-2882

JOACHIMS, Thorsten. *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization*. Pittsburgh, PA : Carnegie-Mellon University. Department of Computer Science, 1996.

KAPLAN, Andreas M.; HAENLEIN, Michael. Users of the world, unite! The challenges and opportunities of Social Media. In: *Business horizons*. 2010. vol. 53, no. 1, p. 59-68. ISSN 0007-6813

KOSINSKI, Michal, et al. Mining big data to extract patterns and predict real-life outcomes. In: *Psychological Methods*, 2016, vol. 21, no. 4, p. 493-506. ISSN 1082-989X

KOTLER, Philip; ARMSTRONG, Gary. *Principles of marketing*. 16th ed. Boston : Pearson Prentice Hall, 2015. ISBN 978-0133795028

LAMBIOTTE, Renaud; KOSINSKI, Michal. Tracking the digital footprints of personality. In: *Proceedings of the IEEE*, 2014. vol. 102, no. 12, p. 1934-1939. ISSN 0018-9219

LAUTERBORN, Robert F. New Marketing Litany: Four Ps Passé: C-Words Take Over. In: *Advertising Age*. 1990. p. 61(41), 26.

LAZER, David, et al. Life in the network: the coming age of computational social science. In: *Science*, 2009. vol. 323 no. 5915, p. 721-723. ISSN 1095-9203

LESKOVEC, Jurij; RAJARAMAN, Anand; ULLMAN, Jeffrey David. *Mining of massive datasets*. 2nd ed. Cambridge : Cambridge University Press, 2015. ISBN 9781107077232

LIDDY, Elizabeth DuRoss. Natural language processing. In: *Encyclopedia of Library and Information Science, 2nd Ed*. New York, NY : Marcel Decker. Inc, 2001. ISBN 0824720717

LOVINS, Julie B. *Development of a stemming algorithm*. Ft. Belvoir, VA : Ft. Belvoir Defense Technical Information Center, 1968.

MAATEN, Laurens van der; HINTON, Geoffrey. Visualizing data using t-SNE. In: *Journal of Machine Learning Research*, 2008. no. 2, p. 2579-2605. ISSN 1532-4435

MAHAPATRA, Lisa. *Social Media Marketing: How Do Top Brands Use Social Platforms?* [online]. September 8, 2013 [Accessed March 17, 2017]. Available from: <http://www.ibtimes.com/social-media-marketing-how-do-top-brands-use-social-platforms-charts-1379457>

MANI, Inderjeet. *Automatic summarization*. 1st ed. Amsterdam : John Benjamins. 2001. ISBN 1588110591

MIHALCEA, Rada; TARAU, Paul. TextRank: Bringing Order into Texts. In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. East Stroudsburg, PA : Association for Computational Linguistics, 2004. ISBN 1932432361

MUSTAFARAJ, Eni; FINN, Samantha; WHITLOCK, Carolyn; METAXAS, Panagiotis T. Vocal minority versus silent majority: Discovering the opionions of the long tail. In: *IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing*. Boston, MA : IEEE, 2011. p. 103-110.

PETERS, Meghan. *Facebook Subscribe Button: What It Means for Each Type of User* [online]. September 15, 2011. [Accessed March 17, 2017]. Available from: <http://mashable.com/2011/09/15/facebook-subscribe-users>

Psychology Today - Home [online]. 2017. [Accessed on May 23, 2017]. Available from: <https://www.facebook.com/psychologytoday/>

ROKACH, Lior; MAIMON, Oded. *Data mining and knowledge discovery handbook.* 2nd. ed. New York : Springer, 2010. ISBN 9780387098227

ROUSSEEUW, Peter J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. In: *Journal of computational and applied mathematics*, 1987. vol. 20, p. 53-65.

SALTON, Gerard.; MCGILL, Michael J. *Introduction to modern information retrieval.* 3rd print. Auckland, N.Z. : McGraw-Hill, 1983. ISBN 0070544840

SANGHVI, Ruchi. *Facebook gets a facelift* [online]. September 5, 2006. [Accessed March 17, 2017]. Available from: <https://www.facebook.com/notes/facebook/facebook-gets-a-facelift/2207967130/>

SPARCK JONES, Karen. A statistical interpretation of term specificity and its application in retrieval. In: *Journal of documentation*, 1972. vol. 28, no. 1, p. 11-21. ISSN 0022-0418

Terms of Service [online]. Date of Last Revision: January 30, 2015. [Accessed March 18, 2017]. Available from: <https://www.facebook.com/terms>

The Nielsen Company. *Social Networks Blogs Now Account for One in Every Four and a Half Minutes Online* [online]. June 15, 2010. [Accessed 17 March 2017]. Available from: <http://www.nielsen.com/us/en/insights/news/2010/social-media-accounts-for-22-percent-of-time-online.html>

XU, Wei; LIU, Xin; GONG, Yihong. Document clustering based on non-negative matrix factorization. In: *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. New York, NY : ACM, 2003. p. 267-273.

ZAFARANI, Reza; ABBASI, Mohammad Ali; LIU, Huan. *Social media mining: an introduction*. 1st ed. Cambridge : Cambridge University Press, 2014. ISBN 9781107018853