Text Mining for Wellbeing: Selecting Stories Using Semantic and Pragmatic Features

Timo Honkela, Zaur Izzatdust, and Krista Lagus

Department of Information and Computer Science
Aalto University School of Science, P.O. Box 15400, FI-00076 Aalto, Finland

Abstract. In this article, we explore an application in an area of research called wellbeing informatics. More specifically, we consider how to build a system that could be used for searching stories that relate to the interest of the user (content relevance), and help the user in his or her developmental process by providing encouragement, useful experiences, or otherwise supportive content (emotive relevance). The first objective is covered through topic modeling applying independent component analysis and the second by using sentiment analysis. We also use style analysis to exclude stories that are inappropriate in style. We discuss linguistic theories and methodological aspects of this area, outline a hybrid methodology that can be used in selecting stories that match both the content and emotive criteria, and present the results of experiments that have been used to validate the approach.

1 Introduction

Wellbeing informatics is an emerging area of research in which ICT methodologies are used to measure, analyze, and promote wellbeing of individuals. Examples of traditional applications include heart rate monitoring, tracking sports activities, analyzing the nutritional content of diets, and analyzing sleeping patterns with mobile technologies. In this article, we present analytical methodology developed for a social media application in which users can find stories that are potentially helpful in their individual life situations. The users may wish to develop their wellbeing further, or need to solve some problem that prevents them from achieving a satisfactory level of wellbeing. The novelty of this paper lies in the overall framework and in its application domain. The specific methods used, such as independent component analysis, are well known as such. Due to the complexity of the overall setting, we evaluate the results in a qualitative manner.

In addition to relying on experts, people have always been listening to advice provided by their family members and trusted friends. This kind of peer support can be very valuable as it is not possible to seek professional advice for all health and wellbeing related questions. Finding a set of suitable stories for a person in a particular situation resembles the basic information retrieval task in which a user is provided with a number of search results.

In the wellbeing case, the information retrieval problem setting is multifaceted. There are at least three criteria that can be used to evaluate stories: a) the match between the problem or situation of the person and the topic or semantic content of a story, b) the style of each story (whether it can be considered appropriate or not), and c) the sentiment of each story (whether the person considers the story encouraging, inspiring, or in general positive or not). These basic evaluation criteria are illustrated in Fig. 1.

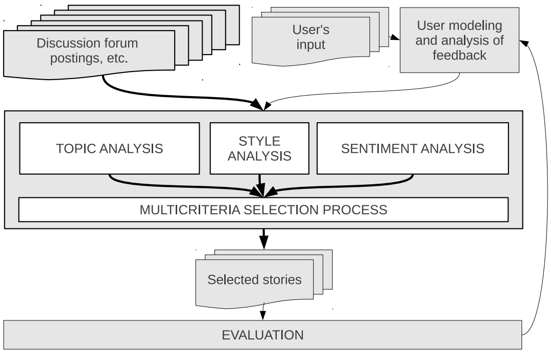
Discussion forum postings, etc.
User's input
User modeling and analysis of feedback
TOPIC ANALYSIS
STYLE ANALYSIS
SENTIMENT ANALYSIS
MULTICRITERIA SELECTION PROCESS
Selected stories
EVALUATION

Fig. 1. The basic architecture of a system that conducts text mining in order to find stories that can support users' wellbeing.

Fig. 1 also provides a wider system context for the present work. The content analysis can be divided into two main areas, i.e., semantic and pragmatic analysis. Linguistic semantics is a research are where computational modeling has traditionally taken place in the framework of symbolic logic. However, adaptive and statistical method are increasingly popular and there are numerous approaches based on neural networks and statistical machine learning. Classical examples include latent semantic analysis [5] and self-organizing semantic maps [14]. In this work, we apply independent component analysis (ICA) in the semantic analysis [2, 8]. This approach is described in detail in the next section.

Whereas semantics predominantly focuses on prototypical meaning, pragmatics is concerned with communicative, contextual and subjective aspects of meaning [7]. From computational point of view, the amount of research on prototypical semantics is much more common than work on pragmatics, mainly due to the efforts invested in knowledge representation and semantic web research. However, there are increasing evidence that the area of computational pragmatics is gaining ground. Research on detection of antisocial behavior from texts [13], and modeling the context of communication [3] can be mentioned as examples. In the context of the present work, analysis of sentiments (see e.g. [15]) and style (see e.g. [12]) are of particular interest.

2 Methods

Here we describe in brief the methods that we later apply for extracting wellbeing-related patterns and features from discussion forum stories. The components of the analysis process, i.e. topic analysis with independent component analysis, style analysis and sentiment analysis are explained in the following sections.

2.1 Topic Modeling with ICA

Independent component analysis (ICA) is a stochastic and an unsupervised learning method for blind source separation. The task in blind source separation is to separate original components (sources) from observed mixtures of random variables without any or with very little a priori knowledge about the components or the nature of the mixing process. In the classic version of the ICA model [11, 4, 10], each observed random variable is represented as a weighted sum of independent random variables. An example of an observed random variable in our case is the frequency of a word in a story.

ICA can be seen as an extension to principal component analysis (PCA) and factor analysis, which underlie latent semantic analysis. ICA is a more powerful technique than these as it is capable of making underlying factors explicit under certain conditions [4, 10, 8]. While PCA is clearly inferior to ICA in terms of identifying the underlying factors, it is useful as a preprocessing technique due to it being able to reduce the dimensionality of the data with minimum mean-squares error. More detailed information on ICA is available in [10] and a detailed presentation on using ICA on text data can be found in [8]. In this work, we use ICA to model and extract topics of stories.

2.2 Sentiment analysis

Sentiment analysis has gained increasing amount of interest as a problem to be solved. It is valuable, for instance, when companies wish to know how the customers are commenting on their services and products in social media and what kind of direct feedback they receive as e-mail messages, through web forms, etc. there are different kinds of methodological alternatives developed for sentiment analysis (see, e.g., [9, 1, 6, 15]).

The basic approach is to build a lexicon where each lexical item (word or phrase) is associated with a negative or positive sentiment value. This lexicon can be built fully by hand, or machine learning techniques can be used to associate values automatically. Linguistic methods are applied for handling constructions such as negation. The overall field of sentiment analysis is too varied to be fully explored here. We rather consider it as one module in the overall system and acknowledge that different choices can be made.

For the sentiment analysis, we used a method called SentiStrength, which has been developed at the University of Wolverhampton, UK[1][15]. SentiStrength

[1] http://sentistrength.wlv.ac.uk/

estimates the strength of positive and negative sentiment in short texts, even for informal language. SentiStrength provides values both for positive and negative sentiments, with scales from -1 (not negative) to -5 (extremely negative), and from 1 (not positive) to 5 (extremely positive). This means that a document can at the same time show both positive and negative sentiments which provides us with more useful information, compared to the regular approach in which only the polarity of a text is determined.

SentiStrength is a lexicon-based classifier that uses negating words, emoticons, spelling correction, punctuation and other kinds of linguistic information in an attempt to achieve a high precision in detecting sentiments[15].

2.3 Style Analysis

Stylistic differences of texts can be most easily defined either by describing the source of the text, or in terms of the genre of the text. Style also deals with the complexity, readability and trustworthiness of texts [12].

In the application context of this work, the most important style factor is the inappropriate use of language such as swearing. When recommendations about potentially useful stories are given, stylistically questionable texts should be excluded. In this work, we applied a simple vocabulary-based approach that is described in the following section.

3 Experiments

3.1 Data and preprocessing

The data used in the experiments consists of stories that have been collected from an internet-based service called Reddit (http://www.reddit.com). Reddit is a social media where registered users provide contents in the form of texts, normally written by the users themselves, or links. The contents cover all aspects of life but we have naturally focused on items that deal with wellbeing. More specifically, selected Reddit channels were “anxiety”, “depression”, “happy”, “off my chest” and “self help”. These include stories that describe both positive (e.g. happiness, accomplishment) and negative experiences (e.g. depression, anxiety).

In the preprocessing phase, punctuation marks were removed and all uppercase letters were replaced by the corresponding lowercase letters. The resulting corpus consists of 2570 documents, 2,886,772 tokens (words in the running text) and 13,023 types (different unique words).

3.2 Independent Component Analysis

In the ICA analysis of the Reddit data, we applied the FastICA algorithm on the word-document matrix. The original dimensionality of the data was first reduced by PCA. Symmetric orthogonalization and *tanh* as the nonlinearity function were used [10].

The vocabulary was manually selected to only cover words that are related to the theme of wellbeing. The full list is too long to be included here but it can be found at http://research.ics.tkk.fi/cog/data/icann12sp/wordlist.txt.

We used the FastICA Matlab package to extract a prespecified number of 20 features. In considering the feature distributions, it is good to keep in mind that the sign of the features is arbitrary. This is due to the ambiguity of the sign: the components can be multiplied by 1 without affecting the model [10, 8].

Examples of the ICA results on the Reddit data are shown in Fig. 2 and Fig. 3. The upper row of diagram in Fig. 2 shows how words "anxiety", "stress", "nausea", and "relief" are associated with the same emergent feature. Similarly, when the lower row in the figure is considered, it is clear that the words "class", "book", "exam" and "professor" have a shared feature. In each case, the representation of the word is quite sparse, i.e., each word is mainly represented by one or two distinguishable features.

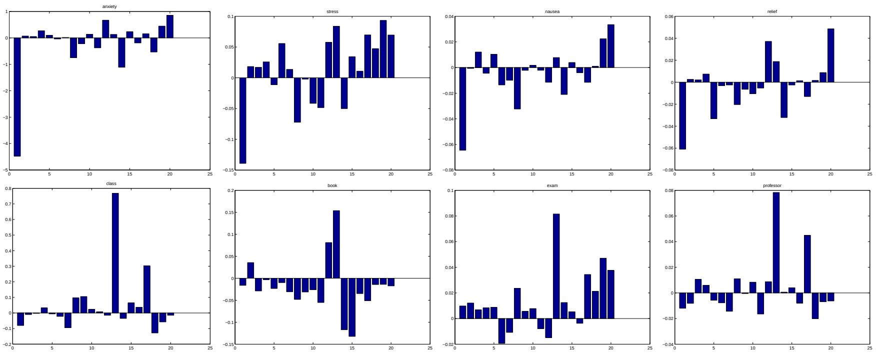

Fig. 2. Upper row: Anxiety, stress, nausea and relief, lower row: Class, book, exam and professor

Another set of result of emergent features based on the use of ICA is presented in Fig. 3. The analysis of the words "problem", "worry", "health" and "motivation" give rise to a a rich representation. These words are clearly associated with several emergent features.

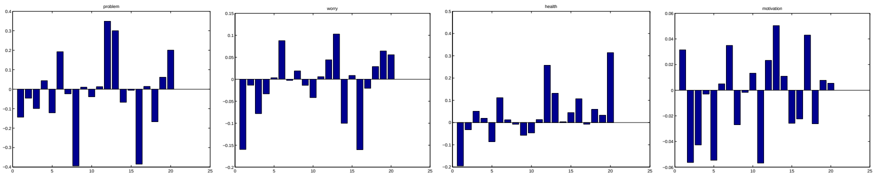

Fig. 3. Problem, worry, health and motivation.

The categorical nature of each emergent feature can be illustrated by listing the words that are strongest in relation to each component. This analysis is presented in Table 1. The order of features is arbitrary but the words most strongly associated with each feature show a clear emergent order. The content of the table is quite self-explanatory but it may be interesting to considered some specific examples. For instance, the feature number 10 is clearly associated with family members and number 8 with alcoholic and non-alcoholic beverages. Feature number 15 appears to collect words that most likely are associated with phenomena that promote wellbeing. This hypothesis receives concrete support in further analysis, discussed later in this paper. If we consider the analysis of the words "health" and "motivation" shown in Fig. 3 and compare them with the features shown in Table 1, we can see that the emergent feature structures clearly make sense. For instance, the word "health" is clearly associated with features numbered 1 ("anxiety", etc.), 12 ("fobia", etc.) and 20 ("doctor", etc.).

In general, it can be concluded that the ICA-based study of a rather small number of documents gives rise to a meaningful representation of the relationship between conceptual items related to the domain of wellbeing.

Feature	Associated words
1	anxiety, nausea, relief, stress, yoga, relaxing (also 19)
2	hate, dread
3	conversation, friends
4	facebook, happy, song
5	chronic, depressed, illness, suffering
6	boyfriend, relationship
7	job
8	alcohol, beer, drink, tea, water
9	friend, girl
10	family, dad, mom, father, mother, parents, brother, sister
11	disability, work
12	fobia, anger, progress, therapist(s), therapy, psychologist (also 20)
13	book (also 15), class, exam, professor (also 17)
14	adrenaline, panic, danger, dying
15	dog, god, love, laugh (also 3), music (also 18), wisdom (also others)
16	worry (also 1), symptom (also others), problem (also 8, 12, 13)
17	school, college, university, homework
18	social, pain
19	sleep, asleep, dream, nightmares, paralysis
20	doctor, drugs, hospital, treat, medical, prescription, google, psychiatrist (also 12)

Table 1. Automatically extracted ICA features.

3.3 Sentiment and style analysis

As mentioned earlier in this article, the SentiStrength method outputs values that assess both positive and negative sentiments in a document. From the analysis, we can see that, overall, the selected subset of Reddit data is biased in the direction of negative sentiments. The results are shown in more detail when they are presented in relation to the results gained with other methods.

The style analysis approach we chose consists of determining the ratio of obscene language to normal language used in the stories. The method simply checks the texts against a dictionary of swear words and then normalizes the results against the length of the corresponding stories.[2].

3.4 Relationships between variables

In this section, the results of two types of analyses of relationships between different variables is presented. First, correlation coefficients between the up and down votes, positive and negative sentiments, the age of the stories, and the ICA components are shown. Assuming that up-vote indicates a preference for reading a story leads us to look at the correlations of that feature with many other features. Regarding correlations with thematic features found by the ICA, the strongest correlation is between ICA features "chronic" and "anxiety" with negative sentiments expressed in the text (see Table 2 for further details).

	Up	Down	Days	Posit.	Negat.	Style
Upvotes	1.0000	**0.7982**	0.0646	**0.2217**	0.0280	0.0099
Downvotes	**0.7982**	1.0000	0.0664	**0.1521**	0.0027	0.0046
Days	0.0646	0.0664	1.0000	0.0255	0.0614	**-0.1245**
Posit. sentiment	**0.2217**	**0.1521**	0.0255	1.0000	-0.0851	0.0456
Negat. sentiment	0.0280	0.0027	0.0614	-0.0851	1.0000	-0.0735
Style	0.0099	0.0046	**-0.1245**	0.0456	-0.0735	1.0000
ICA1: anxiety	0.0462	0.0431	-0.0300	-0.0478	**0.1159**	**0.1138**
ICA4: facebook	-0.3625	-0.2591	-0.0305	**-0.1482**	0.0279	-0.0153
ICA5: chronic	0.0080	0.0235	**0.1296**	-0.0548	**0.1601**	0.0537
ICA9: friend	0.0624	0.0535	-0.0438	**0.1061**	0.0015	0.0805
ICA10: family	**-0.1355**	**-0.1578**	0.0251	**-0.1179**	0.0703	-0.0406
ICA12: fobia	-0.0313	-0.0183	0.0609	**0.1010**	**-0.1184**	-0.0552
ICA14: adrenaline	0.0312	0.0228	-0.0075	-0.0194	**0.1025**	0.0606
ICA15: dog	**-0.1356**	**-0.1047**	-0.0592	-0.2692	0.0167	-0.0155

Table 2. Correlation coefficients between variables.

[2] The list of swear words is available at http://research.ics.aalto.fi/cog/data/icann12sp/

4 Conclusions and discussion

The domain of research presented in this article is novel. It touches upon computational methods, linguistics, psychology and sociology and thus we do not claim to have conclusive results but rather aim to pave way to applications in the area of wellbeing informatics. We see the overall framework as our main contribution. The results highlight the importance of looking at various pragmatic features in addition to semantic ones, when selecting stories that users of wellbeing-related discussion forums find useful.

References

1. Agarwal, A., Bhattacharyya, P.: Sentiment analysis: A new approach for effective use of linguistic knowledge and exploiting similarities in a set of documents to be classified. In: Proc. of the Int. Conf. on NLP (2005)
2. Bingham, E., Kuusisto, J., Lagus, K.: ICA and SOM in text document analysis. In: Proceedings of the 25th ACM SIGIR Conference. pp. 361–362. ACM, New York (2002)
3. Bleys, J., Loetzsch, M., Spranger, M., Steels, L.: The grounded color naming game. In: Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication (2009)
4. Comon, P.: Independent component analysis—a new concept? Signal Processing 36, 287–314 (1994)
5. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. Journal of the American Society of Information Science 41, 391–407 (Jan 1990)
6. Devitt, A., Ahmad, K.: Sentiment analysis in financial news: A cohesion-based approach. In: Proceedings of the Association for Computational Linguistics (ACL). pp. 984–991 (2007)
7. Givón, T.: Mind, code, and context: essays in pragmatics. Lawrence Erlbaum Associates (1989)
8. Honkela, T., Hyvärinen, A., Väyrynen, J.: WordICA - Emergence of linguistic representations for words by independent component analysis. Natural Language Engineering 16(3), 277–308 (2010)
9. Hurst, M., Nigam, K.: Retrieving topical sentiments from online document collections. In: Document Recognition and Retrieval XI. pp. 27–34 (2004)
10. Hyvärinen, A., Karhunen, J., Oja, E.: Independent component analysis, vol. 26. Wiley (2001)
11. Jutten, C., Hérault, J.: Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. Signal Processing 24, 1–10 (1991)
12. Karlgren, J.: Textual stylistic variation: Choices, genres and individuals. In: Structure of Style, pp. 129–142. Springer Verlag (2010)
13. Munezero, M., Kakkonen, T., Montero, C.: Towards automatic detection of antisocial behavior from texts. In: Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011). pp. 20–27 (November 2011)
14. Ritter, H., Kohonen, T.: Self-organizing semantic maps. Biological Cybernetics 61(4), 241–254 (1989)
15. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment in short strength detection informal text. Journal of the American Society for Information Science and Technology 61(12), 2544–2558 (2010)