Identifying Themes in Social Media and Detecting Sentiments

Jayanta Kumar Pal, Abhisek Saha

Global Business Services

HPL-2010-50

Keyword(s):
Social Media, Text mining

Abstract:

Recently, a huge wave of social media has generated significant impact in people's perceptions about technological domains. They are captured in several blogs/forums, where the themes relate to products of several companies. One of the companies can be interested to track them as resources for customer perceptions and detect user sentiments. The keyword-based approaches for identifying such themes fail to give satisfactory level of accuracy. Here, we address the above problems using statistical text-mining of blog entries. The crux of the analysis lies in mining quantitative information from textual entries. Once the relevant blog entries for the company/its competitors are filtered out, the theme identification is performed using a highly accurate novel technique termed as 'Best Separators Algorithm'. Logistic regression coupled with dimension reduction technique (singular value decomposition) is used to identify the tonality of those blogs. The final analysis shows significant improvement in terms of accuracy over popular approaches.

External Posting Date: April 6, 2010 [Fulltext] Approved for External Publication
Internal Posting Date: April 6, 2010 [Fulltext]

Identifying Themes in Social Media and Detecting Sentiments

Dr. Jayanta Kumar Pal
Global Business Services,
Hewlett Packard
jayanta-kumar.pal@hp.com

Abhisek Saha
Global Business Services,
Hewlett Packard
abisek.saha@hp.com

Abstract

Recently, a huge wave of social media has generated significant impact in people's perceptions about technological domains. They are captured in several blogs/forums, where the themes relate to products of several companies. One of the companies can be interested to track them as resources for customer perceptions and detect user sentiments. The keyword-based approaches for identifying such themes fail to give satisfactory level of accuracy.

Here, we address the above problems using statistical text-mining of blog entries. The crux of the analysis lies in mining quantitative information from textual entries. Once the relevant blog entries for the company/its competitors are filtered out, the theme identification is performed using a highly accurate novel technique termed as 'Best Separators Algorithm'. Logistic regression coupled with dimension reduction technique (singular value decomposition) is used to identify the tonality of those blogs. The final analysis shows significant improvement in terms of accuracy over popular approaches.

1. Introduction & objective

In recent years, a significant number of social media, such as blogs/forums turn out to be vast reflectors of customer perceptions. It remains a great challenge to explore this huge information and investigate manually. The difficulty lies in unearthing as much of it as possible in an automated manner through statistical techniques with high accuracy. Further, the textual nature of the data renders them unamenable to regular data analysis.

A company like HP will be interested in tracking these media to identify technical issues faced by the bloggers and ascertain the overall tonality of opinions towards HP and its competitors. In specific, we intend to address the following problems.

- Determine HP's and its competitors' share of voices
- Identify themes/issues related to the business
- Perceive customer sentiments towards HP products and compare with its competitors.

Common techniques extract a fraction of the huge data; screen them manually to come up with solutions to the business questions. In this approach, we address the issues above using several cutting-edge text-mining techniques developed. Those techniques use a small sample of manually rated data as the training set and build learning tools to apply on the overall data, which is much larger in size. This approach significantly reduces the total manual effort and time required to complete the analysis, and makes the solution more robust and accurate.

In next Section we give a brief description of the data followed by illustrations of the method of screening raw textual content to build a concise set of refined words and term-document matrix which capture the relevant information. Section 3 and 4 detail the methodologies developed to address the two core problems along with their implementations, followed by Section 5 describing accuracy results. In Section 6 other commonly used approaches and their relative performances are discussed while Section 7 emphasizes the wide coverage of such application in real life data, followed by concluding Section 8 of references.

2. Data Collection & Preparation

Blogs and comments are extracted from a short-list of technical URL's filtered using some relevant business-specific keywords like printers etc since our focus was on printing industry. The data span Q1 through Q2 2009 and discussions include brands like HP, Canon, Epson, Xerox, Brother and Lexmark. Altogether we had about 850 blog entries in the data-set taken from Australian blog sites.

2.1 Text Parsing & Quantification of Text

Most of blogs/comments are cluttered with words irrelevant to analysis and very much unstructured in conveying information sought for. Therefore Text parsing becomes the first step before any quantitative analysis. First the entire textual content is treated as a long string of characters, which is then scanned and sorted through for a concise set of keywords holding the necessary information. SAS Text miner is employed to execute these steps .Typically it applies the following filters

- Stemmed words like worked/working are treated as the same word;
- Entry category/ Noun-Groups / Product names /Location names or Organization names are considered as different categories;
- Low informative words like articles are deleted;
- Synonyms are grouped together by the same word.

Apart from these there are various other checks that the software does (See [3]) like applying a stop-list where a list of irrelevant word can be pre-specified. After this step we are left with a list of keywords mapped with different documents along with their frequencies in the respective documents. At this point a manual screening was performed just to make the list more aligned with our objectives (i.e. words that have nothing to do with printing/sentiments are deleted). Finally this information is converted into a matrix known as term-document matrix in which frequency of ith keyword in the jth document is displayed at (i,j)th cell of the matrix.. The term-document matrix has been directly used to develop *Best Separators Algorithm* as discussed in next section. The following figure shows the process so far.

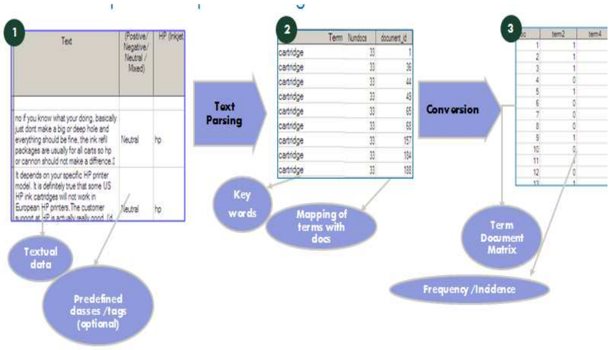
1
Text
(Positive/ Negative/ Neutral / Mixed)
HP (Inkjet
no if you know what your doing, basically just dont make a big or deep hole and everything should be fine, the ink refil packages are usualy for all carts so hp or cannon should not make a diffrence.
Neutral
hp
It depends on your specific HP printer model. It is definitely true that some US HP ink cartridges will not work in European HP printers.The customer support at HP is actually really good. I'd
Neutral
hp
Textual data
Predefined classes /tags (optional)
Text Parsing
2
Term
Numdocs
document_id
cartridge
Key words
Mapping of terms with docs
Conversion
3
term2
term4
Term Document Matrix
Frequency /Incidence

Figure1. Term-document Matrix

2.2 Partitioning the data

We begin by noting that the first problem, question of determining share of voice, translates to scanning each entry string for some specific keywords like HP and its product names, and likewise for its competitors. This is straight-forward and requires no learning tool.

For the next two problems, we split the data in three random parts, training (about 30%), validation (12%) and test (58%) samples. In the next section the approach to theme identification, termed as *Best Separators Algorithm*, is illustrated whereas sentiment detection is addressed in the subsequent section. In both cases training part is used to build the classifier for theme identification and score function which is then validated against validation sample. Finally the validated rules were applied to the test sample.

3. Theme Identification

The problem was to identify broad themes in the document class (one single textual entry henceforth would be called a document and entire textual content together would be called document class). If these broad classes are known beforehand then the exercise becomes a classification problem in which every document needs to be classified as belonging to one or some of the broad themes. Assuming these classes/themes are known beforehand few desirable properties of an ideal classifier are chalked out below, following which *Best Separators Algorithm* is described and shown to have met these properties.

3.1 Multiple classifiers

This classification problem differs from standard classification techniques as implemented in supervised problems in several aspects. Here classes are not mutually exclusive since one document can belong to more than one class. For example, the same blog can talk about a cartridge problem as well as an ink-leakage problem. The solution is to develop separate classifier for each of the themes instead of having one single classifier for all the themes. This makes more sense especially in a scenario where themes are independent from one another, which is true in our case (broad categories of issues are mostly found occurring independently to one another).

3.2 Effect of keywords

Another important aspect where this classification technique significantly claims originality is its feature selection. Most of these techniques use numerical variables as features. Here, whether a document belongs to a specific class is determined by occurrences of few keywords with high **distinguishing ability (**This concept is more precisely and formally defined later when *Best Separators Algorithm* is elaborated). Just to elaborate on this concept, it is quite evident that different words will have different levels of influence in determining the themes. For example, when somebody is pointing to a cartridge issue it is expected that ‘cartridge’ word will have higher weightage over other words like ‘ink’ although ‘ink’ may occur at some of these cases as cartridges involve keeping ink. So influence of ink would be less in these cases. Their roles get reversed when somebody is pointing to ink leakage problem. Motivated by this phenomenon the influential keywords are further categorized in three groups as per their distinguishing ability as principal words, complementary words and auxiliary words.

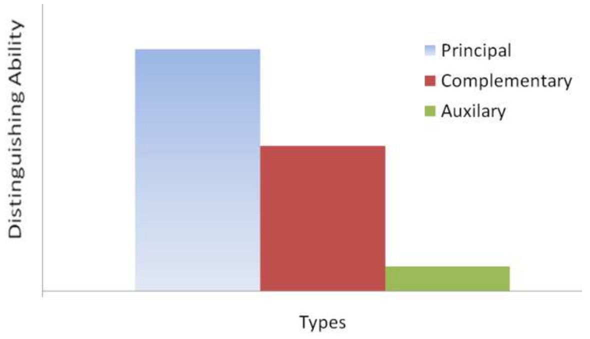

Figure2. Types of influential keywords

So the ideal methodology should consider effects of all these types carefully in building the classification rule.

3.3 Cost of Type I and Type II error

In most of these analyses, various costs could be attached with misclassifications. Especially in a scenario where a classifier assigns an object into one of two mutually exclusive classes misclassification errors become Type I and Type II errors. In our case (which is also pretty common in business) cost of missing an issue (theme) when it is actually present(C(0|1)) is much higher than wrongly classifying a blog as having issue(C(1|0)). This needs to be taken care of by the classifier.

In the next subsection the classifier, termed as *Best Separators Algorithm* is formulated followed by its implementation.

3.4 Best Separators Algorithm

Let the known themes/ classes of issues be denoted here by T_i , i=1 … L**.**

Similarly let W_j , j = 1 … M and D_K , k = 1 … N be the keywords and documents considered for the entire document class as selected in Section 3 respectively.

Let X be the term-document matrix as defined in Section 3. So,

$$X = \left(\left(X_{jk}\right)\right), \{j = 1 \ldots M \ and \ k = 1 \ \ldots \ N\}$$
$$X_{jk} = Frequency\ of\ j^{th}\ keyword$$
$$in\ k^{th}\ document \ldots \ldots \ldots \ldots \ldots \ldots (1)$$

Now,

$$P(W_j \ \mid T_i\ ) = Conditional\ probability\ of$$
$$occurrence\ \ of\ W_j\ given\ T_i = 1.$$

Define

$$DP_i(W_j) = P\left(W_j \left| T_i\right.\right) - P(W_j \left| T_i{}^c\right.),$$
$$\{j = 1 \ldots M \ and \ i = 1 \ldots L\}.$$

Note that $DP_i(W_j)$ represents distinguishing power of W_j in identifying T_i and also note that signed difference is considered .Note that if a high positive value of $DP_i(W_j)$ is observed for some i and j and the document belongs to T_i then there is high chance that W_j would appear in that document. Next, we rank the $W_j{}'s$ in descending order of $DP_i(W_j)$ values for a specific category, say T_i, such that

$$DP_i(Wr_1) \geq DP_i(Wr_2) \geq DP_i(Wr_3) \ldots \geq DP_i(Wr_k)$$

$$where\ \{r_1, r_2, \ldots\}\ is\ a\ permutation\ of\ \{1,2, \ldots\}$$

Consider $W_j{}^{'}s$ such that $DP_i\left(W_j\right) > 0$. Evidently the top-ranked words are the principal and complementary words and the rest would be auxiliary words. The analysis shows a drop in the DP_i values at some threshold after which values would be very close to zero. This is the zone of unnecessary/ non-informative words. Let K_i be point of drop. Define following score function for ith class given a document D.

$$SF_i(D, K_i) = \sum_{j=1}^{K_i} I\left(Wr_j \in D\right) DP_i\left(Wr_j\right) \Delta_j$$
$$\ldots \ldots \ldots \ldots \ldots \ldots (2)$$

$$\Delta_j = \ Premediated\ exponentially\ decreasing$$
$$weight\ assigned\ to\ j^{th}\ ranked\ keyword.$$
$$I(.) \ = \ Indicator\ function.$$

Further the score function assigns exponentially decreasing weights to accentuate the difference in DP_i values. Hence the first few keywords will be playing deciding role in identifying the themes.

$$Say, K \ = \max_{1 \leq i \leq L} K_i$$

Although K_i is chosen observing the drop in the DP_i values for a class i only first few K_s ($\leq K_i$) many words may suffice as it is not straightforward to characterize principal words, complementary words and auxiliary words.

So if we replace K_i by K_s in (1) it is easy to observe that given a document if this score function is very high then there is high chance that this theme is present in the document. Given a D and K_s, a classifier/rule, $R_i(K_s, C_i)$ can be defined as follows.

$$R_i(K_s, C_i) \xrightarrow{identifies\ as} \begin{cases} 1\ (i.e.\ theme\ is\ present) \\ \quad if\ SF_i(D, K_s) \geq\ C_i \\ 0\ (i.e.\ theme\ is\ not\ present) \\ \quad otherwise \end{cases}$$

$$for\ some\ fixed\ constant\ C_i$$

Choosing the best Rule

Borrowing the concepts from Test of Hypothesis two types of accuracies are defined. β = rate of accuracy given the theme is present (similar to power of a Test) and α = rate of error given the theme is not present (similar to significance level of a test) in the training sample.

Since increasing β is more important than minimizing α (see 3.3) best rule is obtained by maximizing β subject to a maximum variation in α of 40% - 50%. Just to create a robust rule α is allowed to be more than 40% only when β is less than 60%.

$$\dots\dots\dots\dots\dots..(3)$$

The best classifier, $R_i{}^*$, would be $R_i(K_s^*, C_i^*)$ obtained by optimizing over all K_s, ($\leq K$) and C_i subject to (3). It can be easily verified that it satisfies all the criteria as laid out in 3.1, 3.2 and 3.3 respectively.

3.5 Implementation

First, the training sample is manually screened to identify 7 broad themes as follows

1. Cartridge related
2. Cost related
3. Color printing related
4. Ink related
5. Ink settings related
6. Printing Media like paper/cloth etc
7. Manual related

For each category a binary variable, say T_i, is specified taking two values as 1/0 based on whether the theme is present or not in a document for training and validation samples respectively. Let INC be the incidence matrix derived from (1).

$$INC = \Big(\big(I(X_{jk} > 0)\big)\Big),$$
$$\{j\ =\ 1 \ldots M\ and\ k = 1\ \ldots N\}$$

INC is combined with T_i's to calculate relative frequencies to estimate $P\big(W_j\ |T_i\big)$ by the following quantity.

$$\sum_k I(INC_{jk}\ = 1|\ T_i\ =\ 1) \Big/ \sum_k I(\ T_i\ =\ 1)\ ,$$

$$k \rightarrow documents\ in\ Training\ Sample$$

Size of the set of influential keywords, K, is found to be 10 by studying the K_i's of the themes. SAS was employed to calculate the $SF_i(D, K_s)$'s for all i's with the Δ_j's taken as normalized principal subset from {0.5,0.25,0.125 ….} of size $K_s(\leq K)$ for all such principal subsets.{ *by principal subset we mean any subset that contains the biggest value and if it contains any other value then it also contains all values in between from the main set*}. Once $K_s{}^*$ and C_i^* are determined the rule is shown to have been working well in validation sample as will be discussed in section 5. Finally we used the rule to predict the themes' presence in test sample.

4. Sentiment Detection

To perform the sentiment analysis, we start with the training set and manually rate these entries as positive / negative/ neutral / mixed. The objective here is to predict one of these classes for each new blog entry.

4.1 Dimension Reduction

Unlike theme identification, these sentiments are reflected by some combinations of keywords, rather than those individually, indicating that we need to exploit the relationship between the keywords. The huge number of words makes it a high-dimension problem and invariably sparsity becomes an issue as number of documents is too less compared to the total number of keywords under consideration. Instead of using frequency directly an entropy weight function on frequency is used as features since it gives importance to very low frequency terms (see [3]). Now a lower dimensional representation of these weighted values of these keywords is obtained employing a cluster preserving Truncated Singular Value Decomposition (TSVD) on the X (as defined in 3.4) matrix through SAS Text-miner (see [3] and [4]).

4.2 Modeling & Implementation

Here, first $n\,(<<\ M)$ SVD components, hereby denoted by $x_i, i = 1\ \ldots\ n$, are taken for analysis. The four mutually exclusive classes, as defined in the beginning, are converted into two binary variables as follows.

1. $Positive = \begin{cases} 1 \; if \; positive/\, mixed \\ 0 \; otherwise \end{cases}$

2. $Negative = \begin{cases} 1 \; if \; negative/\, mixed \\ 0 \; otherwise \end{cases}$

Note that the exercise of predicting the sentiment classes is now equivalent to predicting these two newly defined binary variables as all the four classes can be retrieved back from these two binary variables. We model them by two separate independent logistic models with the lower dimensional components. The models can be specified as follows.

$$Model\ 1\text{:}$$
$$Let\ \pi_1\ be\ Probability\ (Positive\ = 1)$$
$$\pi_1 = e^{(\beta_0+\sum_{i=1}^{n}\beta_i x_i)} \Big/ (1 + e^{(\beta_0+\sum_{i=1}^{n}\beta_i x_i)})$$

$$Model\ 2\text{:}$$
$$Let\ \pi_2\ be\ Probability\ (Negative\ = 1)$$
$$\pi_2 = e^{(\alpha_0+\sum_{i=1}^{n}\alpha_i x_i)} \Big/ (1 + e^{(\alpha_0+\sum_{i=1}^{n}\alpha_i x_i)})$$

We estimated parameters using standard MLE's. Here, the cut-off, n was taken to be 32 since analysis showed they contribute for almost the entire variation in the data. Two logistic models were run in parallel using SAS. Since here also Cost (0|1) is much higher than Cost (1|0) the cut-off was decided to be chosen from 0.10 through 0.33 [≡ C(0|1) = 2C(1|0)]. On grid-optimization over this region the best cut-offs were found to be 0.18 and 0.33 for the mentioned models respectively from training sample. Prediction rule defined based on these cut-offs was implemented in the validation samples as discussed in the next section.

5. Validation/Accuracy Results

Validation sample was also scored and classified manually. It constituted 12% of the entire data-set. Both models are tested to have worked quite well as compared to other known tools in this sample.

5.1 Theme Analysis

For the theme identification analysis, the accuracy rates (fraction of entries properly identified as the theme being present or absent) provide a good measure of its performance. To this end, we use the validation data to measure the accuracy rate. The results are shown in figure 3. We observe that some themes are identified better than some other, due to the sparsity of the latter (the theme itself occurring very rarely in the validation set). The accuracy rates are consistently much higher than 60%, the usual benchmark provided by the usual social media monitoring tools.

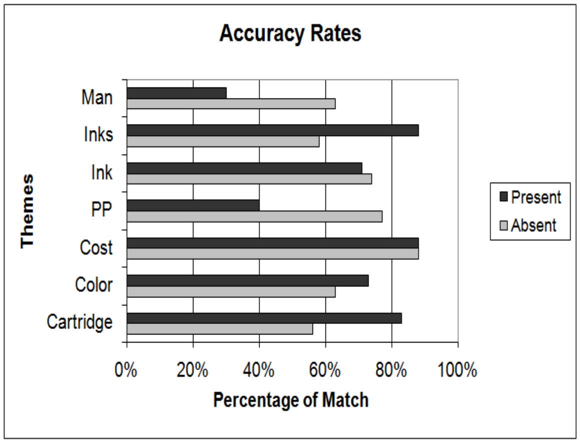

Figure3. Theme Identification

In the above figure the deep black bars show accuracy within class 1, i.e. when the theme/issue is present whereas the light black bars show accuracy within the class 0. The analysis also reveals deep black bars crossed 70% except for two categories, Manual and PP, the reason being these classes were poorly represented in the training data itself as shown in the below table. Again light black bars also remain well above 60% except for two cases.(almost touching 60%).

	Training		Validation	
Themes	No of 0's	No of 1's	No of 0's	No of 1's
Cartridge	178	53	76	24
Color	193	38	89	11
Cost	207	24	91	9
PP	226	5	95	5
Ink	154	77	66	34
Inks	179	52	75	25
Man	219	12	93	7

Table1. Distribution of blogs and comments

5.2 Sentiment Analysis

The sentiment analysis is also validated using misclassification rates, for all the four types (shown in figure 4) except the mixed category, which by nature is difficult to classify due to its rarity and ambiguity by definition, we achieve accuracy rates consistently higher than 60%. Importantly, the positivity and negativity regressions, seen separately, achieve high accuracy rates around 80%. We deduce that the positive and negative sentiments are captured significantly well, and even the combined categories have satisfactory classification rates.

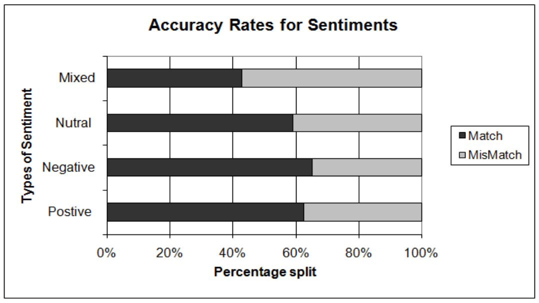

Figure4. Sentiment Detection

The inherent nature of the solution and approach is quite scalable across different Business Units and geographic regions. The solution is being extended beyond specific domains to create an overall platform to track social media. A significant advantage is, for a specific problem area, the manual rating need not be repeated periodically. Once the tool is built using the training data, it will be repeatedly used for future tracking well, in a completely automated manner.

6. Competitive approaches & their relative performances

In context of theme identification there are popular approaches like Naive Bayes(NB) algorithm which uses multiplicative models on occurrences of words in a document. Text classification tools like neural network, vector space models, semi-definite programming (See [1] and [2]), latent semantic indexing(LSI) (See [5]) have been used by researchers in classical text mining problems like spam filtering, document classification etc (See [4]). HP Labs has also developed a tool called Incidence Categorization Analysis based on Naive Bayes algorithm.

The Best Separators algorithm is lot easier to use compared to some of these tools, and does not suffer from volume of keywords and sparsity of the term document matrix. We also observed that the accuracy figures are also higher compared to those commonly used techniques (for both sentiment detection and theme identification accuracy figures hovered in between 50% and 60% for each of NB and LSI when implemented through SAS). The strength of the algorithm lies in the fact that different keywords have different power of classification and uses a weighted data-driven scheme for the learner tool

7. Applicability & future potential

As noted above, the solution by virtue of its scalability readily presents itself as a social media analyzing technique across several business groups. Currently, it is in the process of being advertised, familiarized and adapted to different domains like AR-PR reports, media coverage reports, technical reviews etc. In specific, the techniques are being applied to market reviews from last two quarters to automate its tonality specification in future. Going forward, work is in progress to automate and syndicate the entire end-to-end process starting from extraction of user generated contents to analysis of themes and tonality. The final target is to build a self-supervised tool to capture user feedbacks, both quantitative and qualitative.

Huge sources of textual data, including analyst reports, market reviews, and technical media, have significant impact on the perception of HP in public sentiments. A common theme and sentiment analysis platform will enable HP to identify action areas in all such domains. Influencing them using campaign and correction strategies is of vital importance, and has the potential of creating huge business impact.

8. References

[1] Yiming Yang, Jian Zhang, Bryan Kisiel, "A Scalability Analysis of Classifiers in Text Categorization", *Annual ACM Conference on Research and Development in Information Retrieval, 2003,*96 - 103

[2] Pierre Baldi, Paolo Frasconi, Padhraic Smyth, "*Modeling the Internet and the Web Probabilistic Methods and Algorithms*" Wiley, 2003

[3] SAS Institute Inc. 2004. *Getting Started with SAS® 9 .1 Text Miner*. Cary, NC: SAS Institute Inc.

[4] Michael Berry, Malu Castellanos, "*Survey of Text Mining : Clustering, Classification and Retrieval*", Springer 2008.

[5] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. A., "Indexing by latent semantic analysis" *, Journal of the American Society for Information Science*, 1990, 41(6), 391-407.