SOCIETY FOR SCIENCE AND EDUCATION
UNITED KINGDOM
TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
Volume 8 No 4
ISSN 2054-7390

# Extracting Value from Unstructured Data – Implementing Text Analytics on the Voice of Student

**Jiangping Wang**
*Webster University, USA*
wang@webster.edu

**ABSTRACT**

Unstructured data is chaotic and messy with little or no metadata and lack traditional organization structure. However, same as any structured data, unstructured data is also part of valuable business asset. Many times, it is text heavy and needs extensive preprocessing before data mining algorithm can apply for building models in order to reveal value hidden in the data. Text as a form of data is widely used in business operations as a major way of communication, generating increasing volumes of data. Text data in its raw form is relatively dirty. The embedded business value can be extracted through approaches in text mining and text analytics. This paper presents a case study in this general process of revealing value in unstructured data and applying on data collected to support online learning and student assistance.

**Keywords**: Unstructured data, text mining, text analytics, student learning.

## 1 Introduction

Big data deals with both structured data and unstructured data. Majority of data generated today is unstructured, including data formats in audio, video, image, location, and text. Text as a form of data does not fit into traditional structured data format. Therefore, an extensive preprocessing is necessary in order to reveal value from it. The process of preprocessing unstructured data depends highly on the nature of the data and, many times, domain knowledge is the key to be successful.

Online classes depends heavily on communications over the Internet. Online discussion on topics related to concept that students are learning is an important component. Students are expected to participate discussions actively by researching and learning from discussion questions. In addition, students may ask ad hoc questions through platforms like discussion forum or emails. In the process of online discussion and emailing, data in the form of text is generated. It contains valuable information that reflects students learning, difficulties and concerns they may have. Better understanding student learning and purposefully addressing the issues can significantly improve the quality of teaching and learning. Text mining is a necessary effective approach to achieve this goal. It extracts business value from unstructured text data before extensive text analytics can be further implemented [1].

Most traditional data mining algorithms work well on structured data, including cross-sectional data and time series, but may not directly on unstructured data. Data mining approaches in text mining, social recommendation, and link prediction, are examples of tackling unstructured data, where more efforts in data pre-processing is necessary [2-7].

**DOI:** 10.14738/tmlai.84.8456
**Publication Date:** 26th June, 2020
**URL:** http://dx.doi.org/10.14738/tmlai.84.8456

## 2 Variety in Big Data

In this modern digital age, all forms of business organizations produce large volume of data including both structured and unstructured on a day-to-day basis. Big data features with number of Vs. Although there are different versions of big data definition in terms of number of Vs, variety is always one of the important ones [8-10]. The common acceptable version of four Vs includes volume, velocity, variety, and veracity.

Variety captures the heterogeneous nature of big data [11]. It is a characteristic that defines the variations in the structure of data generated. Big data variety covers all ranges of formats, from traditional structured data to unstructured data. As a matter of fact, the majority portion of volume of big data is collected as unstructured. It grows faster than other type of data. Business value and insights are hidden in the wealth of unstructured data. Unstructured data processing enacts structures on the data as part of retrieval and processing. This flexibility makes it possible to structure the data in accordance with different applications. To unearth the hidden value, additional processing on unstructured data is unavoidable in order to develop needed structure and prepare the data be meaningful for the application and generate value.

## 3 Text Data

Unstructured data is mostly stored in non-relational and schema-on-read database structure for quickly storing any type of data required for analysis with flexibility of having limitless format on any stored data. Traditional data mining methods are expanding their capabilities to take challenges in unstructured data such as text data, social media posts, click stream, geolocation, and sensor data. Unstructured data brings opportunities just same as structured data to enhance the value of data for business in many industries. At the same time, it brings challenges to both business people and IT industry in order to extract values from it to achieve their business goals [12, 13].

Text is everywhere in business operation, yet it is challenging in terms of uncovering value from text form of data. Text as a form data is unstructured. It is generated in every second of business operations and people's social life in documentation, text chat, emails, text streaming, web visit, and all the forms of social media communications. Text is intended for exchanging ideas between people. Therefore, communication context is important, much more so than with other forms of data. The true business value of text data is embedded in bodies of text collections, covered with messy rendering and formatting components that is necessary for human beings to understand. Such components include formatting mechanism in fonts, color, and layout, as well as location URL, email address, and emotional expressions. All of these encoding are noises and need be removed before analytics algorithms can effectively apply to the data. Text mining moves text from unstructured form to the form of a structured matrix before the data can be suitable for standard data mining methods.

The schematic of overall text mining process is shown in Figure 1. It is an iterative process of discovering insights and answering business questions.

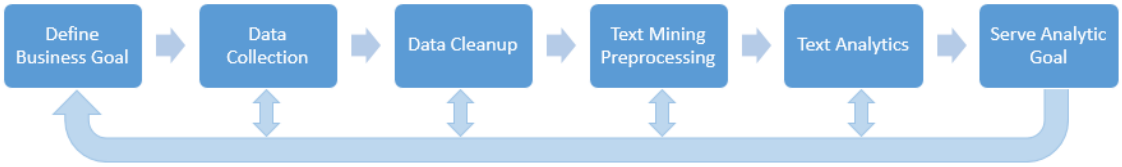



**Figure 1. Overall Text Mining Process**

Copyright © Society for Science and Education United Kingdom

# 4 Case Study

The case study presented below tackles text data collected from online class offerings. The business goal is to better understand how online students perform in classes and identify potential areas for improving student learning.

## 4.1 Data Collection

The operation of any online class offering lacks for direct face-to-face communication between instructor and students. As a result, heavy text communication is a necessary channel in forms of discussion forums and email messages. Discussion forums are created for students posting questions and seeking help from everyone in the class including the instructor. This is an effective way for student sharing their thoughts and experience from each other. Many times a question raised by one student represents the same question from others so the discussion is beneficial to the entire class. Alternatively, students may choose to email their questions and inquiries directly to the instructor seeking private assistance.

For online classes, the discussion section should serve the same purpose as it does in a face-to-face class. The goal of discussion is to engage participation and interaction. The only difference here is that, in a face-to-face class, instructor can purposefully select students, most times targeting those who are relatively quiet at the corner of the classroom and engaging them to respond certain part of the question, which is impossible in an online class. Participation in the process of thinking through questions, reading/learning other students' response, doing research to prepare their own, is more valuable than the answer itself. Students can also use the discussion forum to ask any specific question to seek assistance from others. At the same time, the question and answers from different perspective benefit the entire class. If students do not feel comfortable to share their question or opinion with the rest of the class, they can always email directly to the instructor.

The data for this study was collected from course offerings of upper level data mining class through five semesters in both discussion forum and email exchanges. As expected, some classes generated more discussion/email messages than others did and the length of message varies – some are much longer than others. Therefore, using the data collection from all course offerings for the study is necessary to reflect overall student performance in the course. The data for this study consists total of 165 posts in general ad hoc discussion area and 345 email messages.

## 4.2 Initial Data Cleanup

There are normally a number of preassigned discussion topics in each class to help student master key concepts. In addition, a general question and answer discussion area is established for students asking any question on an ad hoc basis. Questions in this area allow students help each other and answers provided from instructor will benefit entire class. Other students may even have answers to the questions as well from their perspective and experience. Instructor can also use this area to share any discussions originated from private email exchange. A sample screen of student discussion posts is shown in Figure 2.

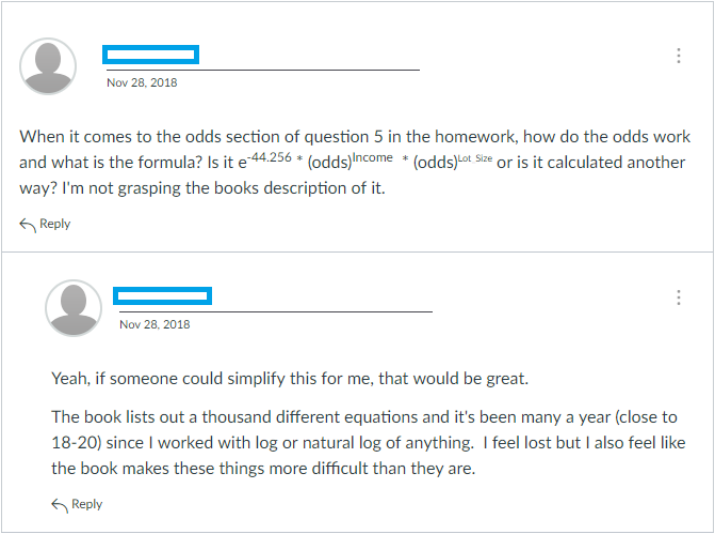



**Figure 2. Sample Student Discussion Posts**

Text data collected for processing include discussion messages from both students and instructor covering various topics, contents, question, answers, and feedbacks. For the purpose of this study, data in the general discussion area is collected. Additionally, questions and responses through class email are also collected. Topics in both areas, discussion and email, reflect issues raised in the course of student learning, ranging from textbook understanding, lab activity, and assignment accomplishment. To protect student privacy contained in the data, any information about student name, class name and class section, date and time are removed in initial data cleanup.

## 4.3 Text Mining Preprocessing

Text, especially those from social media and the Internet, is considered unstructured data. It does not present automatically any feature variables as we see in cross sectional data. Instead, it consists of wealthy linguistic components for human to understand but not for computers. To make computer understand the text data and hence text mining is possible, text preprocessing is a necessary step.

Bag of words is a popular approach to represent free form text with feature variables. Each term, or word, can be a feature variable with value either present or non-present in the body of text, with which frequency of each term can be tallied and meanings can be interpreted in terms of how often each term presents and how important it signifies in a specific text file. More complex preprocessing is required for effective modeling in advanced data mining algorithms on text data.

Going through a series of text mining preprocessing steps, text mining starts turning each word as a feature. Therefore, any formatting component in the text, such as paragraph, extra spaces, and font, should be removed. Some meaningless letters or symbols, such as beta (β), single letters, and numbers, are also removed. Additional text reduction that is necessary includes stopwords removal, stemming, and normalization. The goal is to reveal the feature, which is the collection of meaningful words, embedded in the body of text by eliminating noises of grammar and syntax. This preprocess generates a collection of words that are features and will be further analyzed in various algorithms to serve different analytical goals.

Copyright © Society for Science and Education United Kingdom

At the end, the text mining preprocessing turns the unstructured data into structured data, represented by features in the collection of meaningful words. Table 1 compares the text before and after text mining preprocessing.

## 4.4 Text Analytic Approaches

The ultimate goal in data mining and text mining is to answer business questions. Different approaches can be implemented to serve different goals. In text mining, algorithms can be employed to extract meanings in the documents. In this case, analyzing text data in discussion posts can be on two levels: term level or message level. Term level approaches can be used to better understand meanings embedded in the body of text. By using terms, or words, as units, frequency of each term can be tallied, where several analyzing methods, such as word count, sentiment analysis, and word association, can reveal relationships represented by terms. On the other hand, message level methods are for extracting meaningful pattern carried between pieces of messages. For example, classification model can be built to classify future messages and association algorithm generates association rules among messages.

**Table 1. Text Mining Preprocessing**

| Before Preprocessing | After Preprocessing |
|---|---|
| 16f2w5general.txt - Notepad<br>File Edit Format View Help<br><br>If you stuck on question number 3 for week 1 assignment, here are some extra notes to help you.<br><br>Page 27 of the book states: "to normalize a variable, we subtract the mean from each value and then divide by the standard deviation."<br><br>Therefore, for variable Age (years): Mean =44.66667 and Standard deviation (std) = 14.97554. These values can be easily calculated in Excel functions AVERAGE and STDEV, respectively.<br><br>For normalizing age value for observation #1 (age=25), we calculate as below:<br><br>(25-mean)/std = (25 - 44.66667)/14.97554 = -1.31325.<br><br>In Excel spreadsheet, this calculation can be performed by doing =(B2-B8)/B9, as shown in the | > corpus2[2]<br>[1] " stuck number week assignment extra notes help page book states normalize variable subtract mean value divide standard deviation therefore variable age years mean standard deviation std values easily calculated excel functions average stdev respectively normalizing age value observation age calculate mean std excel spreadsheet calculation performed shown following screenshot normalizing png spreadsheet layout normalized age income values easily calculated calculation follow calculations post sorry sir sent wrong number meant questions please advise tere notes use xlminer categorical dummies data worksheet click xlminer platform transform transform categorical data create dummies create dummies dialogue box move fuel_type right variables factored click ok new worksheet createdummies created last columns data binary dummy variables fuel_type_cng fuel_type_diesel fuel_type_petrol _dummies_ png _dummies_ png _dummies_ png create partitions newly created worksheet named created ummies click xlminer partition standard partition standard data partition dialogue box shows move variables right enable specify percentages bottom panel set partitioning percentages training validation test required click ok new worksheet created named data_partition worksheet click training_data validation_ |

### 4.4.1 Text Meaning Extraction

Word cloud is a popular visual representation in text mining to present meaning in a body of text by giving greater prominence to the words that appear more frequently. Word cloud highlights the most common words and presents the data in a way that everyone can easily understand. An example of word cloud is show in Figure 3, where we can see students generally ask question about data and modeling in weekly assignments. Color in this cloud differentiates frequency categories.

Some of other emphasized terms in the cloud may reflect concerns on learning data mining concepts of building model in such areas as regression, prediction, and odds calculation. If any term in the word cloud is less meaningful for a specific analysis, they can be easily removed in the preprocessing phase before a new word cloud is regenerated.

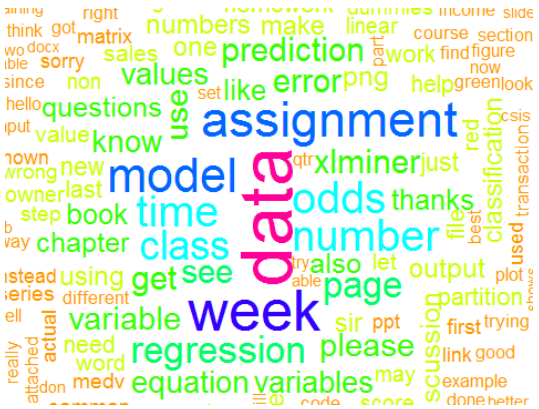



**Figure 3. A Sample Word Cloud**

Cluster analysis helps understand what terms go together, or words were used together in the body of the text, revealing associations. Clustering analysis groups terms into clusters based on distance measurement. Two visual clustering presentations are dendrogram clustering and cluster plot (Figure 4).

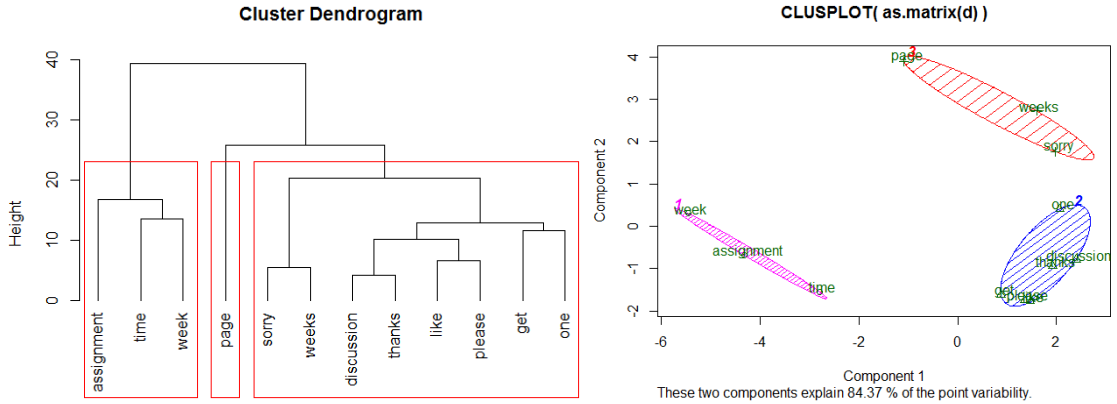



**Figure 4. Dendrogram Clustering and Cluster Plot**

Association in numerical measurement between terms can be calculated easily. For example, analyzing what aspects of assignment students are discussed most often, we understand from the dendrogram that students may have issues and question on assignment time, weekly assignment, and assigned discussion, as shown in Figure 4. The associations with high correlation value, for example over 0.80, are valuable. When a term always appears together with another term, they have correlation value of 1.0.

Sentiment analysis is a form of contextual mining of text that extracts subject information and helps understand students' general opinion towards the class and the topics from the discussion messages. It can be a presentation of making sense of text data in computational score. It is an effective tool to predict topics that students care about and their underlying intention concerning these aspects. It identifies overall attribute expressed in positive, negative, or neutral by assigning weighted sentiment scores to the topics within a sentence or phrase. It certainly helps in our classes to gauge student's opinion, monitoring student learning, understand their education experiences.

Figure 5 shows a histogram of simple sentiment scores on the messages we collected. It represents individual score for each of the ten documents in the collection (17, -20, -8, -3, 3, -4, -16, 5, 5, -5) with average score of -2.6 and standard deviation of 10.80. The Scores were calculated by matching words in

Copyright © Society for Science and Education United Kingdom

each document with a lexicon of positive and negative words [14]. As expected, the overall distribution is slightly skewed and negative on average due to the nature of the text data.

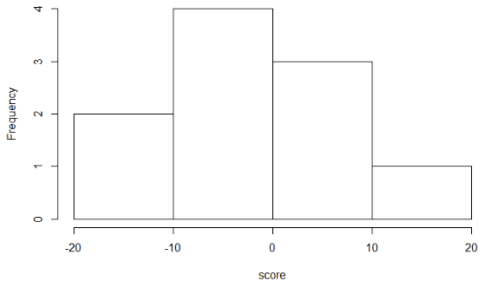



**Figure 5. Histogram of Sentiment Scores**

Term association or correlation can also be performed in text analysis. Association between words indicates what words are used together most often. It is helpful in cases where a term is particularly interested and meaningful to the analysis context. We can identify any words that highly correlate with the specified term. The following example, shown in Figure 6, finds out that the "word assignment" is highly correlated with a few different other words indicating student are interested in assignment time, weekly assignment, and maybe discussion assignment.

### 4.4.2 Message Classification or Clustering

The phase of preprocessing in text mining transforms unstructured data into structured data. The body of text, or corpus of documents can be represented as structured data in the form of term frequency matrix or document frequency matrix, and further matrix measures like TF-IDF (term frequency inversed document frequency). Alternatively, terms may be transformed into concepts as features to represent the text which enhances the manageability of the text for model building purposes. Until then, it is possible to feed the data into many traditional data mining algorithms and build data models for discovering further insights embedded in the data.

```
> findAssocs(dtms, "assignment", corlimit=0.80) # specifying
 a correlation limit of 0.80
$`assignment`
      time       week     please       like discussion
      0.89       0.89       0.88       0.87       0.86
```

**Figure 6. Association between Terms**

The spectrum of data mining covers algorithms in supervised classification, supervised prediction, unsupervised association, and unsupervised clustering. In classification, pieces of message in the discussion forum or email can be classified for different purpose. For example, some messages are for clarification and acknowledgement that do not need immediate response from instructor. To better assist student, instructor can allocate his/her time to respond the message that really need immediate help. To perform this kind of analysis, each piece of collected message needs be coded for the class it belongs for the purpose of model training and validation.

Clustering, as an unsupervised data mining algorithm, form groups or clusters of similar messages. Based on the features that were generated in preprocessing phase, similar messages are clustered in the same group. This helps us better understand students' needs so their questions can be addressed in a better

manner. Clustering at message level seeks relationships among messages whereas clustering at term level focus on the correlation among terms used in messages.

### 4.5 Serving Analytical Goal

Solving business problem is the goal for data analytics. It is important to be clear what business decision to make and what data to analyze. Text mining and text analytic approaches address business issues by targeting the problems that have the most value. Extracting business value from data, including value hidden in unstructured data, can effectively help prediction about future business events and making insightful decisions.

The goal for this text mining project is to better understand the issues in student learning and make improvement based on previous course offerings. As educators, we always seek ways of enhancing the teaching quality. Sometimes, we believe we did good job in instructing students but students may feel differently. Text mining can certainly help us better comprehend how student learn and what issues may exist. Text mining can extract patterns of the common bottleneck and discover the types of key points where students often encounter difficulties in class. It allows us to recognize the areas that require improvement by considering the needs of student, and discover where and how to focus our efforts.

The study in this case can be easily used in other business. The Internet contains a huge amount of text and, for all types of business, understanding customer response often involves understanding text in different ways of communication between customer and the business. The study can be generalized to hundreds or thousands of various types of documents the business is dealing with. This is valuable in the areas of customer service, customer relationship, and product service, where text communication generates large amount of unstructured data. Processing and analyzing these data provide great value to business.

## 5 Conclusion

Text mining is relatively new in the family of data mining techniques. It has been growing fast with more and more data available in plain text format accumulated from software tools such as emailing, text messaging, chat, and social media communication. It is more and more clear that there is abundant valuable business information that can be extracted from text data and used to benefit insightful decision making.

Text mining deals with unstructured data and is more difficult than mining structured data business collected in their operational database, such as data in purchase transactions, customer demographics, and inventory items. Text is unstructured with free word/sentence/paragraph flow representing linguistic sentiments. Therefore, text data need be mining-friendly, meaning preprocessed, before being fed into any mining algorithm and analyzed effectively.

Improvement can be done in the areas such as n-gram sequence, named entity extraction and topic models. N-gram sequence is used to contain sequences of adjacent words as terms in text processing, instead of dealing with individual words only. Named entity extraction recognizes common named entities. Topic model builds a topic layer between document and words to reflect the complexity of language by grouping related words into topics or cluster of words. Domain knowledge is mandatory in these more sophisticated techniques of text processing in order to reach text analytical goal.

Copyright © Society for Science and Education United Kingdom

## REFERENCES

[1] Dang, S. and Ahmad, P. H. 2014. Text Mining: Techniques and Its Application. International Journal of Engineering & Technology Innovations (IJETI), Vol. 1, Issue 4, 22-25.

[2] Rybchak, Z. and Basystiuk, O. 2017. Analysis of Methods and Means of Text Mining. ECONTECHMOD: International Quarterly Journal on Economics of Technology and Modelling Processes, Vol. 6, No. 2, 73–78.

[3] Moreno, A. and Redondo, T. 2016. Text Analytics: the convergence of Big Data and Artificial Intelligence. International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 3, No. 6, 57-64.

[4] Talib, R., Hanif, M. K., Ayesha, S. and Fatima, F. 2016. Text Mining: Techniques, Applications and Issues. International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 7, No. 11, 2016.

[5] Ittoo A., Nguyen, L. M. and Bosch, A. 2016. Text analytics in industry: Challenges, desiderata and trends. Computers in Industry, Vol. 78, 96–107.

[6] Salloum, S. A., Al-Emran, M., Monem, A. A. and Shaalan K. 2017. A Survey of Text Mining in Social Media: Facebook and Twitter Perspectives. Advances in Science, Technology and Engineering Systems Journal, Vol. 2, No. 1, 127-133.

[7] Preethi, B. M. and Radha, P. 2017. A Survey Paper on Text Mining - Techniques, Applications and Issues. IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN: 2278-0661, p-ISSN: 2278-8727, 46-51.

[8] Khan, N., Alsaqer, M., Shah, H., Badsha, G., Abbasi, A. A. and Salehian, S. 2018. The 10 Vs, Issues and Challenges of Big Data. Pro¬ceedings of the 2018 International Conference on Big Data and Education, 52-56.

[9] Leung, C. K. 2015. Big Data Mining Applications and Services. Proceedings of the 2015 International Conference on Big Data Applications and Services, 1-8.

[10] Sinaeepourfard, A., Garcia, J., Masip-Bruin, X. and Marín-Torder, E. 2016. Towards a Comprehensive Data LifeCycle Model for Big Data Environments. 2016 IEEE/ACM 3rd International Conference on Big Data Computing, Applications and Technologies, 100-106.

[11] Kourik, J. L. and Wang, J. 2017. The Intersection of Big Data and the Data Life Cycle: Impact on Data Management. International Journal of Knowledge Engineering (IJKE), Vol. 3, No. 2, 32-36.

[12] Jagadis, H.V., Gehreke, J., Labrinidis, A., Papakonstantinoue, Y., Patel, J. M., Ramakrishnan, R. and Shahabi, C. 2014. Big Data and Its Technical Challenges. Communications of the ACM, Vol. 57, No. 7, 86-94.

[13] Varudharajulu, A. K. and Ma, Y. 2018. Feature-based Restaurant Customer Reviews Process Model using Data Mining. Proceedings of the 2018 International Conference on Computing and Big Data, 32-37.

[14] Hu, M. and Liu, B. 2004. Mining and Summarizing Customer Reviews. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), Seattle, Washington, USA.