Mining Customer Feedbacks for Actionable Intelligence

Lipika Dey, Sk Mirajul Haque, Nidhi Raj
TCS Innovation Labs Delhi
{lipika.dey, skm.haque, nidhi1.r}@tcs.com

Abstract

Mining consumer-generated text can provide business intelligence to organizations by extracting important knowledge trapped in the form of opinions, thoughts, and ideas expressed by their employees and customers on various aspects relevant to business. The key challenge here is to extract and organize relevant information from noisy text in order to effectively transform it to actionable intelligence. However, there is no formal framework to guide this analytical task. In this work, we present a system that is suitable for purpose-driven mining of free-text customer feedback to convert the knowledge gained to actionable intelligence.

1. Introduction

Text mining is gaining rapid grounds as a technology that can provide organizations a more complete view of their business by extracting important knowledge trapped in the form of opinions, thoughts, and ideas about products, services, competitors or any other allied entities in unstructured text sources. Unstructured text sources may include emails and blogs and discussion forums, free-text feedbacks obtained through direct customer survey, customer call records, analyst reports, news sites or any other relevant source. Given that the data is generated in an uncontrolled fashion, the key challenge here is to extract relevant information and interpret it in the context of business requirements. Though Natural Language Processing (NLP) based analytics like phrase extraction etc. can aid the analysts in this task to some extent, these tools do not perform well on noisy text data due to poor grammar, and spelling errors. Pure dictionary-based analysis mechanisms are also not completely reliable for analyzing noisy text data due to spelling errors and non-standard vocabulary.

In this work, we present a text mining framework for multi-perspective analysis of free-text customer feedbacks. The proposed system allows analysts to determine regions-of-interest within the data and then drill-down in a systematic way. We demonstrate that an amalgamation of natural language processing techniques with machine learning mechanisms can yield better actionable intelligence.

The rest of the paper is organized as follows. Section 2 provides an overview of related work. Section 3 presents the overall architecture of the proposed system. Section 4 presents the details of the text mining components used. Section 5 presents an approach that can be adopted for systematic analysis of feedbacks. Section 6 presents some results on different text repositories.

2. Review of related work

Text mining is now a relatively mature area of research. It is being increasingly applied to trap the knowledge stored within unstructured repository. In [1], a generic web text mining lifecycle built on top of GATE [2] was presented.

An important aspect of knowledge discovery from customer feedbacks involves opinion mining. After [3] proposed a list of positive and negative opinion words, several supervised and unsupervised learning approaches have been proposed to learn opinion expressions from text [4-5]. Approaches to identify product feature mentions and their associated opinions were described in [6-9]. A large number of commercial systems also exist in the market. Though some of them use machine learning and NLP concepts, these systems are predominantly word-based and do not produce accurate view of the content. Document clustering is a standard technique for grouping text to get consolidated views. The fuzzy c-means (FCM) algorithm [10] has been utilized in a wide variety of engineering and scientific disciplines. Semi-supervised clustering techniques can make use of initial knowledge provided to the system to produce clusters in a controlled way. One such implementation is presented in [11]. The proposed framework differs from the earlier systems since it allows the user to perform focused analysis.

978-0-7695-4191-4/10 $26.00 © 2010 IEEE
DOI 10.1109/WI-IAT.2010.196

3. Overview of the proposed text mining platform

The proposed text mining platform performs natural-language processing driven statistical processing of text. Consumer generated text originating from multiple heterogeneous sources are aggregated into a single repository before they are processed. The text processing life-cycle consists of the following steps:

Step 1: *Cleaning and preprocessing of the feedback text*: Detailed description of the preprocessing steps can be found in [13].

Step 2: *Natural Language Processing:* The main objective of this module is to identify linguistic components of a sentence using NLP tools like tagger, parser and dependency analyzer. We have employed Stanford Parser[1] to extract all noun, verb and adjective phrases of length four or less without considering the stop words for analysis. The d*ependency extractor* extracts the dependency relationships between a pair of words in a sentence. These are used to identify subject-object, subject-action, action-object, and subject-feature relationships.

Step 3: *Text Mining:* During this phase, the linguistic and semantic components are grouped together, evaluated and interpreted as quantifiable measures. Details of the text mining platform are presented in the next section.

4. The text mining platform

In this section, we propose an ontology-based text mining system that exploits domain ontology to provide insight into business-critical information in a multi-faceted way. Ontology encodes business intelligence in terms of relevant entities, actions, services or service attributes that are relevant for business. The task of feedback analysis is to identify identifies these elements and their associated opinions in feedbacks. The aim is to provide an aggregated multi-perspective view of the information.

4.1 Ontology based feedback analysis

Though ontology building is an application-dependent task, Figure 1 presents a generic ontology applicable across domains. The set of concepts here are organized under different conceptual heads like products, services, departments, problems etc. Entity descriptions consist of property names, values, functionalities and any other known attributes. The problem concept stores possible problem categories along with expected problem descriptions, associated entities, causes etc.

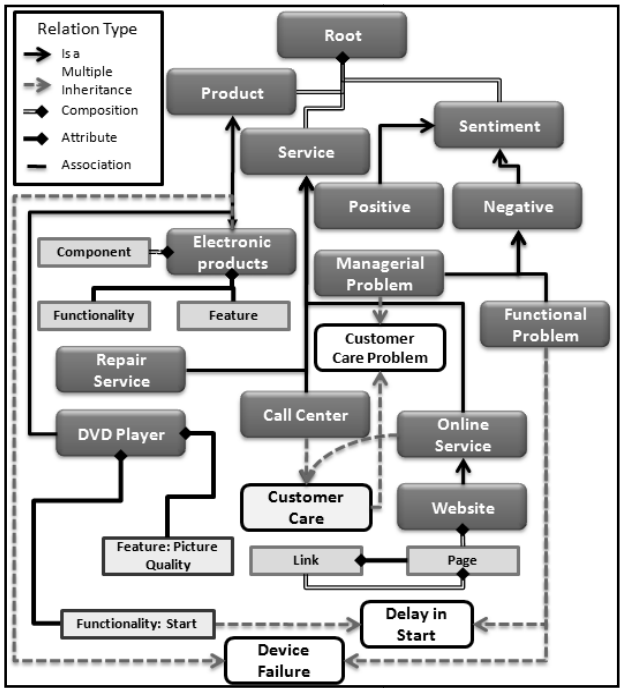

Figure 1: Partial view of an organizational ontology

The ontology structure serves as a basis for aggregating feedbacks under different categories. All explicit ontology concepts within feedback texts are identified through word or phrase-based matching taking into account the Parts-of-Speech of the words and their synonyms. The type of a phrase is also taken into account.

A feedback sentence is assumed to contain a subject of discussion which is an entity or event, one or more of their associated actions or properties and also some opinion words or indicators of positive or negative opinions. The following activities pave the way for systematic analysis of feedbacks using the affect analysis framework.

Subject or entity extraction: The noun entities which have associated descriptions or opinions are identified as potential subjects of feedback.

Action or event extraction: Actions or events are interpreted entity-action relationships through verbs and verb-phrases in sentences.

Descriptor extraction: Functionalities, features and opinions associated to a subject of discussion are interpreted by analyzing the dependent verb phrases, adjective phrases and other noun phrases using a dependency graph that is built from the dependency relations output by the parser.

[1] http://nlp.stanford.edu/software/lex-parser.shtml

Phrase normalization and phrase grouping: Semantically similar phrases are grouped together as follows:

(a). *Phrase normalization*: Phrases with same set of key words (in their root forms) are grouped together irrespective of the ordering of the words.

(b). *Overlap-based phrase grouping*: A wholly subsumed phrase is merged with a larger phrase, if the sentences referred to by them have a large overlap.

(c). *Synonym based phrase grouping*: Two phrases which contain different but synonymous key words are grouped together.

Opinion Mining: An *opinion expression* is a group of words extracted from the text that comprise of opinion words, subject of opinion and also modifiers of opinion, if any. Modifiers are words that can alter the strength or orientation of an opinion word. The base opinion mining algorithm presented in [13], was enhanced to consider negative sentiments in a generic way without dependency on keywords.

4.2 Semi-supervised fuzzy clustering to aggregate noisy text

The NLP based methods described earlier cannot identify all the phrases in noisy text. Assuming that the earlier step can provide an initial support for identifying different domain concepts in the text the next step is to apply semi-supervised clustering mechanisms to identify similar but noisy concepts from the text. Initial supervision to the clustering process is provided in the form of phrases and words representing regions of interest. Thereafter a modified fuzzy clustering algorithm is employed to group similar feedbacks together.

5. Deriving business insight from feedbacks

Support of a concept in the feedback repository is the total number of feedbacks that contain the concept. Relevance of a feedback to the concept is computed from its membership value to the cluster that contains this node. The support for lower level concept nodes in the ontology nodes is propagated to the higher level nodes using algorithm proposed in [12]. Each node also accumulates cumulative positive and negative score based on opinions present in the member feedbacks. Ontology-guided repeated drill-down can provide a complete picture in a systematic way. Figure 2 shows a sample sequence of activities for product-related reviews posted on Amazon after a new TV-DVD player combo product was launched. The ontology concept *dvd-player* is selected for gaining further insight. It is found that the associated action "take" has high negative sentiments associated to it in conjunction with *start*. Drilling down further, it is discovered through the descriptive phrases that it is due to the delay to start. Figure 2 also shows aggregated opinion scores for both *dvd-player* and its *start* functionality.

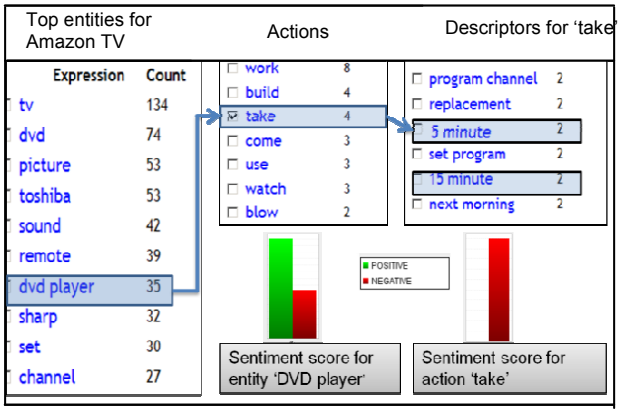

Figure 2. Mining activity to gather customer feedback on TV-dvd combo from Amazon

6. Implementation and results

The system has been *implem*ented as a web-based application interacting with a back-end database, in which the original feedbacks along with all analytical components extracted from them are stored. We present here results from three case studies. The first repository contains data collected from the discussion forum of a leading health food company. The second repository contains feedback data collected from multiple web sources for a services company. The third repository comprises of feedbacks collected through a web-based customer survey for a product. Table 1 shows the average accuracy of entity-action or entity-description modeling tasks executed over different repositories.

Table 1: Average accuracy of text mining tasks

Type	Recall	Precis-ion	F-measure
Find entities for a given action or description	0.7676	0.8636	0.8128
Find actions associated to entities	0.9753	0.8943	0.9330
Find descriptions for given entities	0.8095	1	0.8947

Table 2 presents average accuracy of the clustering process over the repositories. It can be seen that the average gain in support through this process is significant in all cases.

Table 2. Accuracy of fuzzy clustering process

Repository	1	2	3
Total number of feedbacks	3,000	10,000	2,000
Feed-backs retrieved by top 20 phrases initially	840	2033	715
New feedbacks further added through clustering	1346	4649	1227
Gain (%)	160	228	171
Accuracy of clustering(%)	82.27	85.39	75.45

Table 3 shows the precision and recall for the fuzzy clustering process when feedbacks are grouped into three clusters, where each cluster aggregates feedbacks from a different perspective. The three perspectives used here are products, services/business goals and generic sentiments. The first group contains reviews of a known product. The second cluster contains information about satisfaction with intended actions. The third cluster grouped all sentences with sentiments either about the above or other elements like company, competitors etc. Figure 3 shows the relative overlap of the comments when viewed from different perspectives. Little overlap among the clusters show that all the three perspectives are essential to get a complete insight.

Table 3. Accuracy of multi-perspective clustering for health food feedbacks

	Precision	Recall
Product	0.940291	0.835074
Goal	0.79021	0.620879
Sentiment	0.580087	0.642994

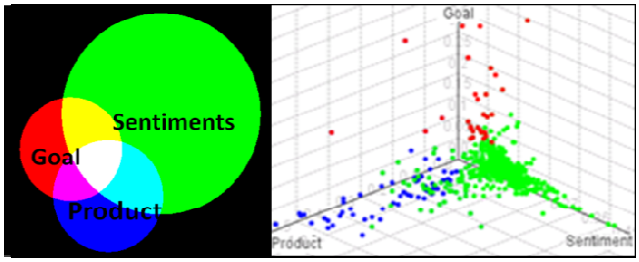

Figure 3. Depicting overlap and separation of clusters for Health Food Company feedbacks

6. Conclusion

In this paper we have presented ontology-based text mining framework that can analyze customer feedbacks in a systematic way. The system is equipped with multiple, efficient text-mining components which can extract high quality relevant information from feedback texts and organize them in an effective way to provide accurate contextual insight of text. Presently the system is being extended to act as a continuous feedback analysis platform with predictive capabilities to generate early warnings to avert undue negativity.

7. References

[1] Castellano, M., Mastronardi, G., Aprile, A. and Tarricone, G. A Web Text Mining Flexible Architecture, World Academy of Science, Engineering and Technology 32 2007

[2] GATE – General Architetcture for Text Engineering, http://gate.ac.uk/

[3] Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings* of the OpinionWeight 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL, pages 174-181, Madrid, Spain, July.

[4] Jeonghee Yi, Tetsuya Nasukawa. 2004. "Sentiment Analysis: Capturing Favorability Using Natural Language Processing" In *KCAP* 03, October 23-25, Florida, USA.

[5] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment Classification Using Machine Learning Techniques. *EMNLP'2002*, 2002.

[6] M. Hu and B. Liu. Mining and summarizing customer reviews. *KDD'04*, 2004.

[7] A-M. Popescu and O. Etzioni. Extracting Product Features and Opinions from Reviews. *EMNLP-05*, 2005.

[8] N. Jindal, and B. Liu. Mining Comparative Sentences and Relations. In *AAAI'06*, 2006.

[9] Ding, X., Liu, B. and Yu, P.S. A Holistic Lexicon-Based Approach to Opinion Mining. WSDM'08, February 11-12, 2008, Palo Alto, California, USA

[10] Richard J. Hathaway and James C. Bezdek. Fuzzy c-Means Clustering of Incomplete Data. SMCB, 31(5) :735--744, 2001.

[11] Huaxiang Zhang, Jing Lu, Semi-supervised fuzzy clustering: A kernel-based approach, Knowledge-Based Systems ,Volume 22, Issue 6, August 2009, Pages 477-481

[12] Lipika Dey, Shailendra Singh, Romi Rai, Saurabh Gupta: Ontology Aided Query Expansion for Retrieving Relevant Texts. AWIC 2005: 126-132

[13] Lipika Dey, Mirajul Haque, S.K, Opinion mining from noisy text data, IJDAR 12(3), 205-226, (2009).