Sentiment Classification of Drug Reviews
Using a Rule-Based Linguistic Approach

Jin-Cheon Na, Wai Yan Min Kyaing, Christopher S.G. Khoo, Schubert Foo,
Yun-Ke Chang, and Yin-Leng Theng

Wee Kim Wee School of Communication and Information
Nanyang Technological University
31 Nanyang Link, Singapore 637718
{tjcna,kymwai,assgkhoo,assfoo,ykchang,tyltheng}@ntu.edu.sg

Abstract. Clause-level sentiment classification algorithm is developed and
applied to drug reviews on a discussion forum. The algorithm adopts a pure
linguistic approach of computing the sentiment of a clause from the prior
sentiment scores assigned to individual words, taking into consideration the
grammatical dependency structure of the clause using the sentiment analysis
rules. MetaMap, a medical resource tool, is used to identify various disease
terms in the review documents to utilize domain knowledge for sentiment
classification. Experiment results with 1,000 clauses show the effectiveness of
the proposed approach, and it performed significantly better than baseline
machine learning approaches. Various challenging issues were identified
through error analysis, and we will continue improving our linguistic algorithm.

Keywords: Sentiment Classification, Drug Reviews, a Rule-Based Linguistic
Approach.

1 Introduction

With the explosion of Web 2.0 platforms, there are enormous amounts of user-
generated content, called social media. Therefore, for the past decade, many
researchers have been studying effective algorithms for sentiment analysis (or
sentiment classification) of user-generated content [7]. Sentiment analysis is a type of
subjectivity analysis which analyzes sentiment in a given textual unit with the
objective of understanding the sentiment polarities (i.e. positive, negative, or neutral)
of the opinions toward various aspects of a subject. It is still considered as a very
challenging problem since user generated content is described in various and complex
ways using natural language. Digital libraries are about new ways of dealing with
knowledge, and researchers are considering the problem of organizing and searching
digital objects, not just by standard metadata fields but also by sentiment polarities [8].

For sentiment analysis, most of researchers have worked on general domains (such
as electronic products, movies, and restaurants reviews), but not much on health and
medical domains. Previous studies have shown that this health-related user-generated
content is useful from different points of view. Firstly, users are often looking for
stories from “patients like them” on the Internet, which they cannot always find among their friends and family [11]. Moreover, studies investigating the impact of social media on patients have shown that for some diseases and health problems, online community support can have a positive effect [6]. Because of its novelty as well as quality and trustworthiness issues, user-generated content of social media in health and medical domains is underexploited. Therefore, the objective of this paper is to develop an effective method for sentiment analysis of social media content in health and medical domains. The sentiment analysis is applied to drug reviews on a discussion forum. In the following sections, related work is discussed first. Then our proposed sentiment analysis method and its experiment results are presented and discussed. Finally, conclusion is provided.

2 Related Work

Researchers have used various approaches for sentiment analysis [7]. Most of the early studies were focused on document-level analysis for assigning the sentiment orientation of a document [9]. However, these document-level sentiment analysis approaches are less effective when review texts are rather short, and in-depth sentiment analysis of review texts is required. More recently researchers have carried out sentence-level sentiment analysis to examine and extract opinions toward various aspects of a reviewed subject [4]. In contrast to most studies which focused on document-level or sentence-level sentiment analysis, our approach uses clause-level sentiment analysis so that different opinions on multiple aspects expressed in a sentence can be processed separately in each clause. For instance, the sentence “I like this drug, but it causes me some drowsiness” has two clauses expressing two aspects: overall opinion and side effects. Some researchers have studied phrase-level contextual sentiment analysis, but phrases are often not long enough to contain both sentiment and feature terms together for detailed analysis [13].

Generally there are two main approaches for sentiment analysis: a machine learning approach (or a statistical text mining approach) and a linguistic approach (or a natural language processing approach). Since clauses are quite short and do not contain many subjective words, the machine learning approach generally suffer from data sparseness problem. Also the machine learning approach cannot handle complex grammatical relations between words in a clause. Some researchers used various linguistic features in addition to the bag-of-word (BOW) feature in the machine learning approach to overcome the limitation of the BOW approach [14], which is being a popular trend in sentence-level sentiment analysis. In this study, we are using a pure linguistic approach to overcome these weaknesses of the machine learning approach. The main advantages of a pure linguistic approach are that we can define sophisticated rules to handle various grammatical relations between words in a sentence or clause, and new rules based on linguistics can be incrementally added to the system. Our previous work performed sentiment analysis at the clause-level using a linguistic approach and focused on movie reviews [12]. This study is based on the previous work, and we are improving the approach by adding additional rules for handling more complex relations between words and adapting it to a new domain, drug reviews.