EXTRACTING OPINION FROM UNSTRUCTURED DATA

Atul Antil

Undergraduate student, University of California, Sandiego, US

ABSTRACT

To separate and decipher general supposition from the casual portrayal of content in online life sites. Techniques: The casual portrayals containing suppositions are Tokenized, Parts of Speech Tagging, Word-sense Disambiguation, and Text Transformation/Attribute Generation are utilized. Test information relating to execution rating of cricket players was gathered from twitter, Cricinfo and cricbuzz. The audits were appraised against emotional assessment criteria, the scale going from poor, moderate, great, phenomenal and the phonetic factors were changed over into numerical qualities utilizing fluffy since the communicated sentiments might be straightforward. Conclusions upgrade any basic leadership process, so the impact of feelings from test information relating to execution rating of cricket players is delegated poor, moderate, great and brilliant. Utilizing the proposed methodology, a nonexclusive model can be assembled that can be utilized to remove feelings and decipher.

1. INTRODUCTION

Emotions and opinions are the two important logical substance of human life. This in turn produces large data in the social media sites. Opinions can be used to extract valuable information that influences decision making1. This article portrays how raw text containing opinions drawn from social media sites are preprocessed, categorized, summarized, evaluated and the influence is furnished to the decision makers. Text preprocessing involves text cleanup, tokenization, part of speech tagging and attribute generation. Text categorization classifies the document collection of many possible user-dependent and application-dependent categories/classes. These categories are generated with distinct characteristics of the opinion to be evaluated. Summarization involves Sentiment classification which uses polarity assignment to express the opinions with positive assessment, negative assessment or neutral assessment. The rest of the paper is organized as follows, Section 2 deals with some existing work on opinion mining, Section 3 discusses the importance of opinion mining and Section 4 presents the Text mining approach to extract opinions from unstructured text. Experimental results are discussed in Section 5 and finally Section 6 concludes the paper by giving the glance and the future direction of research in this area.

2. STATE OF ART

The data produced from blogs, discussion forums, social media and social networking sites, reviews are in the form of unstructured text. In2 suggested text mining techniques are the way to pre-process and extract information from the unstructured text to make better business decision making.

In3 presents that identity, sharing, conversations, reputation, groups, relationships and presence are the seven building blocks of social media. In4 opinioned that corporate makes use of social media that offers abundant occasions for gathering user preferences, opinions, ratings about a product or service and assessments.

In5 suggested that the sentiments attached with a product, the perception about a brand and the perception about new product introduction can be well interpreted using sentiment analysis. In6 depicted an alternative way of classifying opinions like exciting, irrelevant and objectionable, rather than positive, negative and neutral.

According to7 sentiment analysis with the help of fuzzy logic deals with reasoning and gives closer views to the exact sentiment values. According to8 an opinion mining system can be used for both binary and fine-grained sentiment classifications of user reviews. Feature-based sentiment classification is a multistep process that involves preprocessing to remove noise, extraction of features and corresponding descriptors and tagging their property using fuzzy functions.

3. OPINION MINING - ASSESSMENT OF PEOPLE

Generally speaking, opinion mining or sentiment analysis aims to determine the attitude of the user with respect to some topic or the overall contextual polarity in social media data analysis. The attitude may be his or her judgment or evaluation, affective state, or the intended emotional communication. Natural language processing, text analysis and computational linguistics are used collectively to mine opinions and extract information pertaining to a subject contained in the input source9. The focus of sentiment analysis is to classify the polarity of the input text as positive or negative or neutral, at various levels like document, sentence or feature. Sentiments are classified based on emotions expressed in the sentence like “angry”, “sad” or “happy”.

3.1 Approaches to Opinion Mining

Traditional opinion mining approaches can be divided into four main categories:

- Keyword spotting is an approach that classifies text based on the affect words like boring, sad, happy or afraid10.
- Lexical affinity detects affect words and in addition it also assigns the “affinity” to a given emotion11.
- Machine learning methods like support vector machines and Naive Bayes classifier contributes to Statistical Methods12.
- Concept level approaches were based on knowledge representation techniques like ontology and semantic networks13,14.

To extract the context sensitive meaning of a given opinion, the grammatical relationship between words are taken into consideration15.

4. TEXT MINING APPROACH TO EXTRACT OPINIONS

The main sources of data are from newswires, reports, official websites related to a particular topic, microblogs which are unstructured or semi-structured generally. The data found may vary in formats. Extracting data and analysing the extracted data is a time-consuming process and this leads to poor decision making. Classical text mining techniques are applied for extracting information and converting the information to structured data using the suitable categorization methods16.

Figure 1 shows the application of text mining methods to extract user opinions from social media data with a step-by-step process which includes collecting/gathering, pre-processing, information extraction, text mining operations, evaluation/interpretation and finally decision making. Text Pre-processing involves Noise Removal, Tokenization, Parts of Speech Tagging, Word-sense Disambiguation and Text Transformation/Attribute Generation. “Bag of Words” and “Vector Space Model” are Text representation methods based on feature words and their occurrences, where every word is represented as individual variable with a numeric weight attached to it. Feature Generation and Feature Selection methods based on features contained in a document are used for Information Extraction17.

The proposed approach makes use of a domain specific feature knowledge base guide the feature extraction process. The fine-grained selection of keywords can be achieved by defining a property P that finds the keywords from the vocabulary of the text that are mostly adjectives.

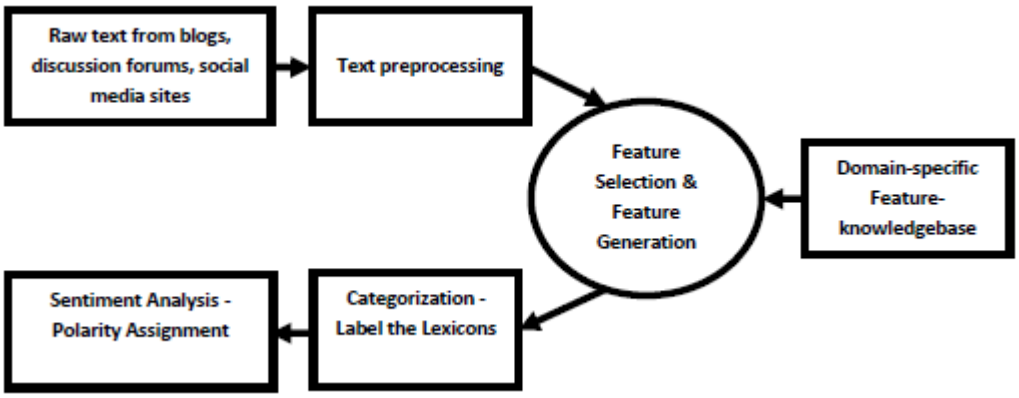

Figure 1. Text mining approach to extract opinions from unstructured text.

P(w) is True, if and only if w is from the vocabulary of adjectives V

The mathematical representation of fine-grained selection of keywords Ks is given by:

$$K_s = \{w \mid w \varepsilon V \text{ and } P(w)\}$$

Interpretation/Evaluation - The final step in text mining is Evaluation/Interpretation that brings about either termination or iteration. When the desired results are achieved, the text mining process can be terminated else further iterations can be made to accomplish the required results. The results can be

visualized and reports can be presented in a way that is easy for the user to understand using Business Intelligence tools18,19.

5. EXPERIMENTAL RESULTS

Sample data was collected from twitter, cricinfo, cricbuzz. They were pertaining to performance ratings of cricket players. Since data were collected from various sources, they were mostly unstructured. The input data was pre-processed to remove stopwords like exclamatory marks, question marks, emoticons, smilies contained in reviews of the users. Python programming was used to carry out pre-processing. Figure 2 shows the sample screen of the pre-processed data. The first part shows the sentence-level annotations and the second part shows the word-level annotations.

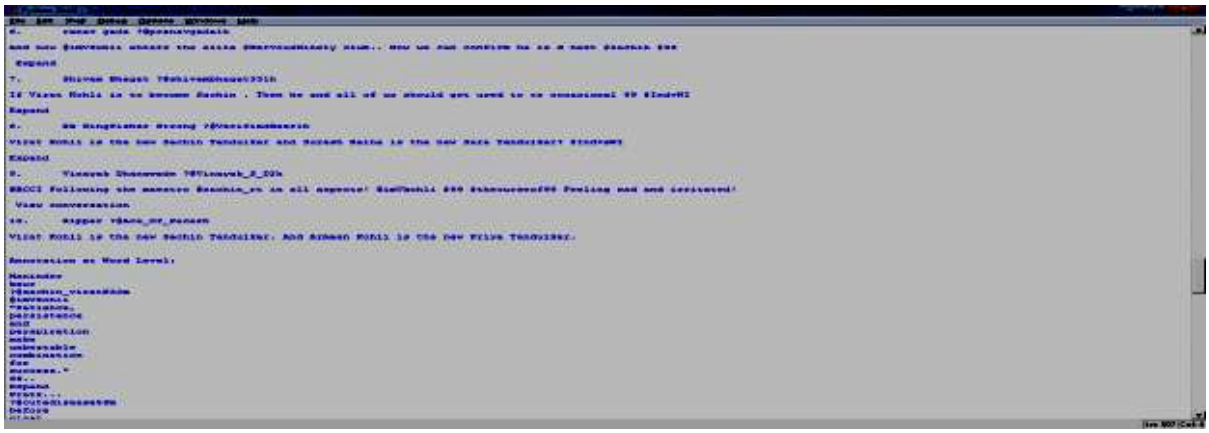

Figure 2. Pre-processed data after stop word removal and tokenization.

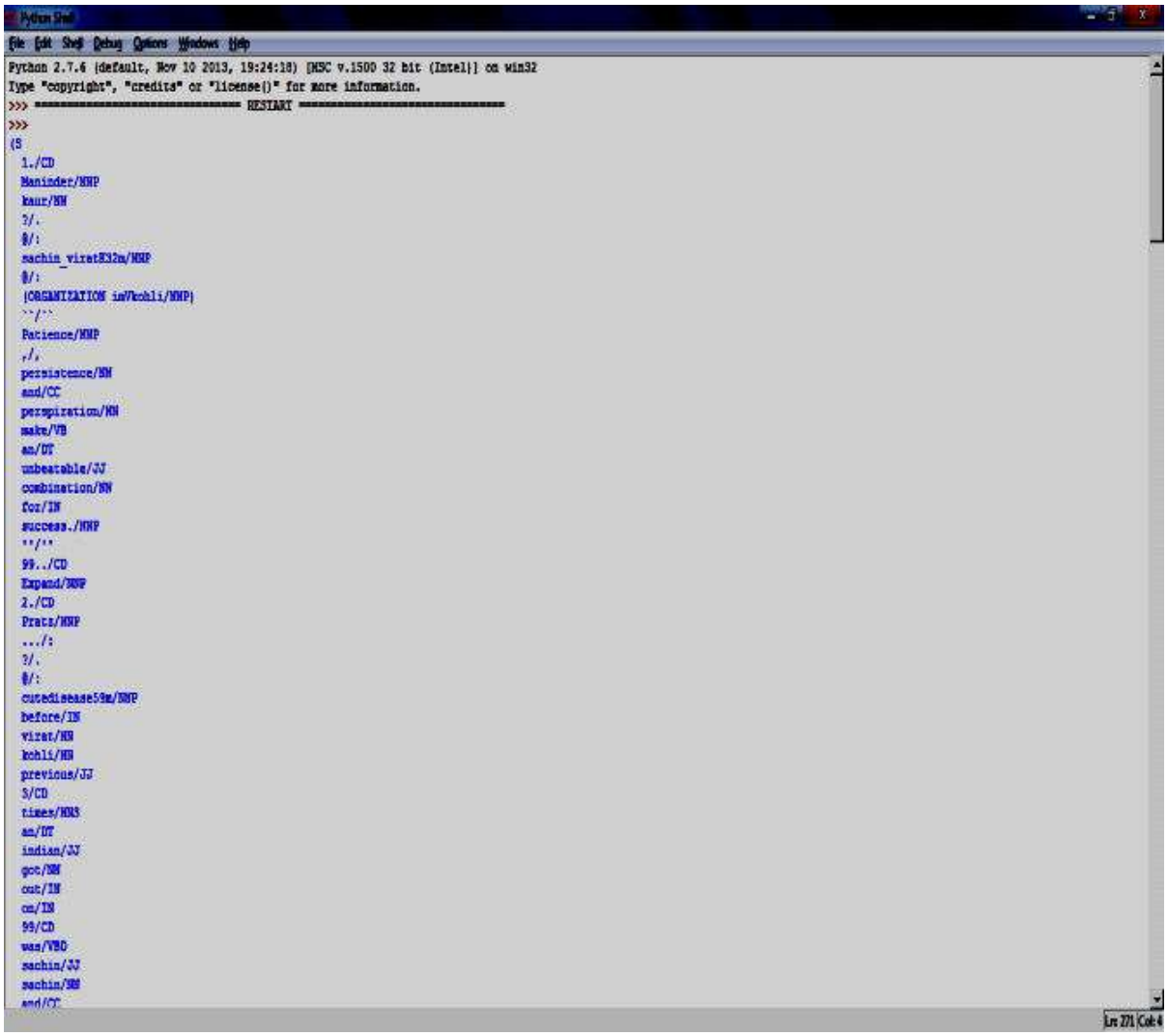

Figure 3. Screenshot of POS tagging using Python and NLTK.

The pre-processed data was subjected to Part-of-Speech tagging with the focus to identify nouns, verbs and adjectives in order to interpret the opinions or comments collected. Python and Natural Language ToolKit(NLTK) were used for this purpose. Figure 3 shows the output of POS tagging.

The separated reviews were appraised against emotional assessment criteria with help from WORDNET, the scale running from poor, moderate, great, amazing and they were spoken to utilizing the fluffy set. By changing the semantic factors into numerical qualities utilizing fluffy, the communicated sentiments might be spoken to in a structure that is anything but difficult to understand20.

F(opinion) = {0.2,0.4,0.6,0.8}

Where 0.2 denotes poor, 0.4 denotes moderate, 0.6 denotes good and 0.8 denotes excellent.

During pre-processing, extra effort had to be applied to establish uniformity of syntax and interpreting the expressed opinion. 6. Conclusion

Opinion mining or sentiment analysis is a special field of Text analysis. Text Mining is more improved than data mining since a deeper understanding of the business domain is attained thereby contributes in deriving competitive business intelligence from unstructured data sources. This article aims to extract valuable information influencing business decisions from raw text. Sentiment analysis is really a challenging area of research for different reviewers. The future work is to fine tune the model and make it generic so that when any opinion is given as input the influence of the opinion on decision making is measured and interpreted.