ISSN 2250-3153

A Natural Language Processing based Web Mining System for Social Media Analysis

John Selvadurai

PhD Student at Indiana State University

Abstract- Social Media Monitoring and Analysis are the new trends in technology business. The challenge is to extract correct information from free-form texts of social media communication. Natural Language Processing methods are sometimes used in social media monitoring to improve accuracy in extracting information. This paper discusses a web mining system that is based on Natural Language Processing to analyze social media information. In that process, this research examines Natural Language methods that are important for such analysis. Then the traditional web mining steps are discussed along with proposed use of Natural Language Processing methods.

Index Terms- Natural Language Processing, Social Media Analysis, Web Mining, Text Mining.

I. Introduction

The advent of social media has changed traditional business methods drastically. Product development, marketing and support are increasingly conducted via social sites. Social Media allows individuals around the world to freely express their opinions without disclosing too much identity. This allows customers to provide genuine opinions about products and services in social media sites. Many business organizations interact with their customers and colleagues through social media.

As the businesses increasingly use social media, the need for better understanding of the trends and communication arises. Social media monitoring is a growing technology area in which sophisticated tools are used to monitor social interactions towards certain text terms. Applications of this technology are huge. Stock market analysis, Political campaigns, Crime monitoring, and Product research are a few of them. Competitive Intelligence is another application of social media monitoring that is receiving a lot of momentum recently. Competitive Intelligence is systematic process of gathering information about an organization's external environment.

The challenge in social media monitoring is to correctly interpret user communication. Usually, the communication entered by users in social media is unstructured free-form text. That means the data does not follow a predefined model. Because of this free-form text nature, simply searching for keywords in social media communications is not an adequate method. Certain sophisticated text mining methods are used in some systems to fulfill the need for accurate social media monitoring. Some other social media monitoring systems use Natural Language Processing methods to extract correct information. This paper discusses a system that extensively uses Natural Language Processing and Web Mining techniques to analyze social media communications.

II. Natural Language Processing

Natural Language Processing (NLP) is a study in Computer Science that converts human languages into computer languages. In other words, NLP allows computers to understand human language such as English in a meaningful way. There are numerous NLP based applications that exist currently. Spelling and grammar check in word processing applications are one of the examples of NLP applications. In modern systems, NLP approaches are combined with statistical models to derive better results. NLP has many sub areas that are facilitating the interpretation of human languages. This paper discusses few NLP approaches that are essential for social media monitoring.

A. Automatic Summarization

Automatic Summarization is a technique that reduces a large text into a meaningful short paragraph. This approach retrieves the summary of large text documents. Internet search engines use this technique to identify the content of a website so that can be indexed appropriately.

In general, there are two types of approaches used in Automatic Summarization. One is Extraction which extracts few important parts from the text such as key word sentences or paragraphs [1]. Another type of approach is Abstraction which paraphrases the important points of the text. Technically, Abstraction is a more complicated system than Extraction to develop. Mainly, each Auto Summarization system's functionality can be divided into three general steps [1].

1) Analysis:

In the analysis stage, the text is analyzed and an internal representation is generated. This internal representation is produced in the way logical relationships between sentences can be established.

2) Transformation:

In Transformation stage, the internal representation is manipulated to produce an ordered text representation. In general, the ordered representation could have sentences ranked by a scoring function. The scoring method is usually based on the facts found in analysis stage. A simple example is where sentences with frequent keywords get a specific score.

3) Realization:

In Realization stage, the summary is generated based on the scoring of Transformation stage. In some systems it could be simply producing specific scored sentences from the text.

There are many auto summarizing tools available. smmry.com is an online tool that produces summaries either by providing URL of a web page or uploading a file. smmry.com also provides API for web developers who could use auto summarization for their developments [2].

B. Named Entity Recognition

Named Entity Recognition is the subject area that identifies entities or physical objects such as names of persons, organizations, and places in the text. For an example, Named Entity recognition of sports news would consist of names of the players, places of teams and grounds, etc. Traditionally, grammar based approaches are used to identify name and entities in the text. In the present systems, statistical models are incorporated in order to classify names and entities more precisely. In statistical approaches, initially, a set of training data will be used against the model. Based on this training data, statistics will be prepared. These statistics will be used against real documents [3]. The advantage of statistical approach is it can be continuously improved as it is used on various texts. Historical data patterns will be used in identifying new name entities.

The Natural Language Processing Group at Stanford University offers variety of solutions to NLP problems. Stanford NER is a Named Entity Recognizer which is implemented in Java platform and provides libraries with a way to identify names and entities [4].

C. Part-of-speech tagging

In this NLP approach, sentences will be tagged according to the grammatical order such as nouns, verbs and adjectives. Because of the real nature of a language, some words can have multiple tags such as both noun and verb. Part-of-Speech (POS) tagger programs use combinations of several techniques such as lexicons, rules, and dictionaries [5]. Dictionaries contain categories of words. Usually, tagging programs accurately tag the word or make a best guess. When the words are ambiguous in a sentence, POS taggers use probability approaches to tag correctly.

D. Word-sense Disambiguation

Word-sense Disambiguation is a NLP subject that identifies the correct sense of the word in a sentence where that word could have multiple meanings. This area is an important part in information retrieval because the contextual meaning of the text depends on the correct interpretation of that text. Human languages generally have many ambiguities. A human can understand the meaning of a word based on the context it is spoken and the background knowledge of the subject. However, a machine would have difficulties in identifying such correct meaning. Word-sense disambiguation methods help a machine to minimize the ambiguities of words in the text.

In general, a word-sense disambiguation approach consists of a lexical repository that contains different senses for words [6]. WordNet is a free lexical database in English that contains a large collection of words and senses [7]. The design of WordNet was inspired by the theories of human linguistic memory [8]. WordNet encloses a large volume of nouns, verbs, adjectives and adverbs in English language. In WordNet, words are grouped and interlinked by their meanings. This method allows the identification of any closer or similar meaning to a given word. Many NLP applications use WordNet as a source for processing.

E. Sentiment Analysis

Sentiment Analysis is an NLP process which identifies the attitude or contextual polarity of the writer with respect to the text. A text collection could show one or many sentiments. Sentiments can be positive, negative or neutral. Sentiment Analysis is heavily used in processing online reviews for products, movies and books. Many websites run specific algorithms to rate the reviews based on the degree of positive or negative tone.

In general, an opinion consists of two components; a target and a sentiment towards the target [9]. Usually, sentiment analysis is conducted after the text is parsed by Part-of-Speech tagging. Specific rules will be applied along with an internal dictionary to POS tags to identify the sentiment. A simple example of a sentiment rule is if two adjoining positive words are adverb and adjective then it is classified as positive sentiment. There are more sophisticated algorithms that exist in order to extract the sentiment.

III. Web Mining

Web Mining is the technique used to extract useful information from data gathered from internet. The term mining is used to express the idea of extracting valuable substance from raw material.

Traditional Data Mining techniques are used to gather information from large scale data bases called data warehouses. Web Mining uses those Data Mining techniques to extract information from internet. This paper discusses a few Data/Text Mining approaches that could be used in Web Mining to build an intelligent system that would analyze social media information.

Web Mining can be classified into three major types; Web Usage mining, Web structure mining and Web content mining. Web usage mining is the process of capturing web user behaviors. This usage related information allows understanding the effectiveness of web sites. Generally, web usage mining involves mining the web server logs to find the user login or visits information. Web Structure mining process identifies relationship between web pages and their links. This tool helps search engines to identify and classify web pages by information. Web content mining directly deals with the content of the web. Even though, search engines attempt to provide precise content, still the information retrieved is missing high accuracy.

In general, the text mining approach consists of four steps [10]:

A. Data Gathering

In this first step, necessary data will be collected from social media. Initially, this data is considered as raw data because no valuable information is extracted at this point. Almost all the major social media sites offer API access to external developers. In industry specific analysis such as stock market analysis, only

specific keywords can be used to identify required data. The challenge is to gather historical data because many social media sites provide access to only current data up to certain number of days. Also, storing large volume of historical data is costly for many application developers. In these cases, social data providers who store historical data from social sites and licensed to resell data to application developers. GNIP is such a social data provider who resells data from major social media sites [11].

In traditional web mining, gathering data is just searching and storing. Size of social data available is so huge because millions of users around the world enter data continuously. Gathering the entire data from social media sites is costlier in terms of storage space. Therefore, this paper suggests using NLP Auto Summarizing technique to reduce the size of gathered data. Since auto summarizing reduces the size without losing the main points, this web mining tool should produce the required condensed data.

B. Preprocessing

In preprocessing step, the raw data will be processed to provide a platform for data analysis. The main purpose of this step is to classify raw sentences into a machine readable form. Generally, this machine readable form is an Attribute-Value table. In an Attribute-Value table, text and its characteristics are identified as model and its attributes. The entire document or page to be parsed will be represented as Attribute-Value table. Since social data is entered by users as free-form text, classifying the data into a table is a challenging task. In order to achieve this purpose, certain Natural Language Processing techniques such as Part-of-speech tagging and Named Entity Recognition will be used. This approach is rather expensive in terms of storage needed for entire attributes of characteristics of the document. However, a carefully parsed Attribute-Value table is essential for precise interpretation. This paper proposes to use Word-sense Disambiguation along with other NLP techniques to increase the accuracy of the resulting data classification.

C. Indexing

Preprocessed texts are indexed and stored in this step. The efficiency and performance of the program depends on the type of indices chosen in this step. Current search engines do this very efficiently. When choosing the appropriate index method, storage types, look up speeds and fault tolerance must be considered. These factors are different for each industry and organization. Cost of storage and speed of retrieval are always a concern in any commercial application. In many situations, in order to reduce storage cost few indexes are created. Even though this approach reduces the cost, it increases the possibility of fault occurrences. However, in some industries, fault tolerance cannot be acceptable. In those situations, indexing must be carefully implemented at the expense of storage cost.

Inverted index is the most common index method used in text retrievals. In inverted indexing, lists are made from terms that appear in the text collection [12]. A simple example of inverted indexing is the index published at the end of books where each word and page numbers are given.

D. Mining

This is the part where actual information will be extracted from the indexed classified data. Frequent Pattern Mining is one of the common tasks occurs in this step.

Frequent Pattern Mining is a method used to analyze frequent behaviors of persons or entities. This method was first introduced to analyze customer buying behaviors from retail transaction databases [13]. In Frequent Pattern Mining, a behavior is identified as frequent by simply finding the count of that behavior. For an example, a simple frequent pattern mining is finding milk buying pattern of customer is to count the number of times milk appeared in the transactions of that customer. Finding how many times a new movie is discussed in Twitter is another example. More complicated patterns of mining exist to find voting and stock performance behaviors. Another application consists of opinion mining which to find opinions about a product or service in social media. Finding the user opinion about a movie in social media is a simple example. This can be done by using sentiment analysis to determine how positively or negatively the movie is discussed in Twitter.

IV. Conclusion

This paper discussed a web mining system for social media that uses NLP techniques frequently. As part of the discussion, specific NLP techniques are listed in detail. Each of these NLP techniques can be subjected to further research. Also, this paper discussed steps in traditional web mining and proposed methods to incorporate NLP techniques for better results. Traditional web/text mining techniques are not popularly used in social media monitoring. The main reasons for this lack of interest are the free-form style and the huge volume of social media text. In this paper, NLP techniques are combined with traditional web mining techniques to suggest a social media monitoring system. This theoretical system is a base concept and many applications could be implemented from it.

References

[1] D.Bikel and I.Zitouni, Multilingual Natural Language Processing Applications: From Theory to Practice, USA: IBM Press, 2012, pp. 400.

[2] A. Elmaani, (2012, Nov 27). SMMRY [Online]. Available at http://smmry.com/api

[3] D.Bikel and I.Zitouni, Multilingual Natural Language Processing Applications: From Theory to Practice, USA: IBM Press, 2012, pp. 286.

[4] The Stanford Natural Language Processing Group. (2012, Nov 15). Stanford Named Entity Recognizer (NER) [Online]. Available: http://nlp.stanford.edu/software/CRF-NER.shtml#About

[5] Robin, (2012, Nov 12). Natural Language Processing: Parts-of-speech tagging, [Online]. Available: http://language.worldofcomputing.net/pos-tagging/parts-of-speech-tagging.html

[6] S. Bandyopadhyay, S. Naskar and A. Ekbal. Emerging Applications of Natural Language processing, Hershey, PA, USA: IGI Global, 2013.

[7] Princeton University. (2012, Nov 10). WordNet: A Lexical database for English [Online]. Available: http://wordnet.princeton.edu/

[8] M. Song and Y. Wu, Handbook of Research on Text and Web Mining Technology, Hershey, PA, USA: IGI Global, 2009, pp. 194.

[9] B. Liu, Sentiment Analysis and Opinion Mining, USA: Morgan and Claypool, 2012.

[10] H. Prado and E. Ferneda, Emerging Technologies of Text Mining: Techniques and Application, Hershey, PA, USA: IGI Global, 2009, pp. 56.

[11] GNIP. (2012, Nov 28). GNIP: The Social Media API [Online], Available: http://gnip.com/sources

[12] J. Lin and C. Dyer, Data-Intensive Text Processing with MapReduce, USA: Morgan and Claypool, 2010.

[13] M. Song and Y. Wu, Handbook of Research on Text and Web Mining Technology, Hershey, PA, USA: IGI Global, 2009, pp. 228.

AUTHORS

First Author – John Selvadurai, PhD Student at Indiana State University, MBA, M.S. Computer Science, B.S. Computer Science. sjohnandrew@gmail.com.