International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:7, No:4, 2013

Opinion Mining Framework in the Education Domain

A. M. H. Elyasir, K. S. M. Anbananthen

***Abstract*—**The internet is growing larger and becoming the most popular platform for the people to share their opinion in different interests. We choose the education domain specifically comparing some Malaysian universities against each other. This comparison produces benchmark based on different criteria shared by the online users in various online resources including Twitter, Facebook and web pages. The comparison is accomplished using opinion mining framework to extract, process the unstructured text and classify the result to positive, negative or neutral (polarity). Hence, we divide our framework to three main stages; opinion collection (extraction), unstructured text processing and polarity classification. The extraction stage includes web crawling, HTML parsing, Sentence segmentation for punctuation classification, Part of Speech (POS) tagging, the second stage processes the unstructured text with stemming and stop words removal and finally prepare the raw text for classification using Named Entity Recognition (NER). Last phase is to classify the polarity and present overall result for the comparison among the Malaysian universities. The final result is useful for those who are interested to study in Malaysia, in which our final output declares clear winners based on the public opinions all over the web.

***Keywords*—**Entity Recognition, Education Domain, Opinion Mining, Unstructured Text.

I. Introduction

THE growing rate of the internet with all services provided made the education domain more reliable and stronger. Top schools, universities and institutions supported various methods to evaluate their facilities, staff and services to enhance and make better environment for students. Online feedback is the most common method to obtain opinions and comments from the majority, then analyze, categorize the data accordingly, tabulate and provide a meaningful evaluation result. Majority of the feedbacks are unstructured and requires sophisticated methods to extract the desired information from this undefined format of text. The unstructured property, no pre-defined format and has no place in relational tables, of the feedbacks makes the advancement in data mining field necessary and required sophisticated method to extract the unstructured text.

Opinion mining or sentiment analysis is a new area within the field of text mining, arose as a developed technology to provide better method in analyzing the unstructured text. It concentrates not simply on the act of retrieving information from relevant documents and sentences, but on extracting information about the overall sentiment and opinions contained within the text. The framework consists mainly of three main stages: extraction, text processing and polarity classification. The chosen extraction techniques vary from information retrieval (IR), natural language processing (NLP) to machine learning techniques. For optimum word extraction, several sub-stages are included for pre-processing and tagging such as focused web crawling to search and collect the opinions from various online resources, HTML parsing to remove HTML tags and frames as well as indexing and storage, a segmenter to prepare the data for feature extraction. POS tagging concludes the first opinion mining stage in the framework by determining which parts of speeches to be assigned to a particular word.

We start the next stage with Stemming and Stop Words Removal to eliminate high frequency words. Stemming is to transform the convertible words back to its root such as the word "walking" is transformed to "walk" to maintain one standard for latter text classification. Within the same framework stage, we employ Named Entity Recognition (NER) to extract all relevant information for latter analysis, this process is extremely necessary in our framework because our domain is undiscovered and no specific tagged corpus exist for the education application for various reasons.

Universities globally compared against each other with benchmarks that specify particular characteristics under international standards. Those standards are of academic and official concerns, i.e. researches and publications, number of graduates and the quality of their future work, university prestige and honor depicted in the held events and ceremonies, etc. However, public benchmarks where everybody shares his/her voice regarding particulars are rarely found, therefore we utilize the internet which is the biggest platform for the public to share their opinions and compare some popular Malaysian universities against each other (locally).

Classification is the last stage in our framework. The sentence polarity is determined during this stage with clear indication of the polarity's degree using Positive, Negative or Neutral keywords based on numerical score. Generally, there are two classification types, sentiment sentence classification and sentiment document classification. The sentence classification applies when the polarity of sentence level is crucial while classification on document level matters for overall and general polarity. In our framework, we tend to use sentiment sentence classification by utilizing SentiWordNet [2] to provide us with the word polarity and classify the overall result accordingly.

The framework details will be discussed deeper in the

Ayoub Mohamed H. Elyasir is with the Faculty of Information Science and Technology, Multimedia University, 75450 Melaka, Malaysia (email: ayoub_it@msn.com).

Kalaiarasi Sonai Muthu Anbananthen is with the Faculty of Information Science and Technology, Multimedia University, 75450Melaka, Malaysia, (phone: 606-2523344; fax: 606-2318840; e-mail: kalaiarasi@mmu.edu.my).

contribution section, where more explanatory diagrams are provided to discover the flow of the framework. A brief research on opinion mining background is introduced in the related work.

II. Related Work

Opinion mining is widely used nowadays in different domains and applications for decision making process; politics utilize it in the elections to find major votes and candidates' popularity in the internet, large companies evaluate their own products by looking into online users' opinions and feedbacks, analyze human behavior and psychology through the social networking sites using sentiment analysis tools and build classifiers to understand the complexity of human natural language.

Reference [3] recently proposed a study on the commenters' behavior and their inter-relationship with their counterparts from comment platform, by developing a sentiment patterns to detect the characteristics of the behavior and relationships in the political domain.

Opinion mining plays a significant role in the marketing where the companies are concerned about the user reviews on particular products or services. Reference [1] conducted extensive study using the Amazon data sales and reviews about digital cameras to get the notion of the customer preference on certain features of the product; they also attempt to predict the sales changes and pricing using the same textual data.

Psychology and emotion analysis benefited a lot from opinion mining applications in social networking sites, whereby some e-commerce enterprises are interested in knowing the consumer behavior and their interests, needs, purchases habits and culture. For example, collection of random samples of comments from the online was classified to positive and negative emotions to understand the comment background and driving emotion behind it, the study has proved that MySpace is rich environment for emotional comments [4].

Opinion mining is a powerful tool nowadays, especially when it is applied on the internet and its social network platforms. Most of the successful businesses utilize enterprise sentiment analysis systems like Clarabridge to find out the customer requirements and help them in the decision making process.

III. Our Contribution

We propose and demonstrate a comprehensive opinion mining framework to search for opinions and comments on various Malaysian Universities, prepare and process the gathered opinions for polarity classification and then evaluate each university according to its polarity result. To process the gathered opinions in a systematical manner and make our framework pragmatic and practical, we divided it into three main stages: Extraction, Pre-processing and Classification. Fig. 1 shows the beginning of the sentiment analysis, in which opinion gathering from various resources in the internet takes place using our focused web crawler. We packed different functionalities in our focused web crawling to do searching and retrieving, parsing and indexing to prepare the opinions for punctuation classification. The searching module is important for finding the popular pages which is seen by most of the users and open up the APIs for information retrieval in the education domain. Indexing module can help in providing information about the web pages ranking that gives the crawler the ability to collect the pages based on their respective ranking and prioritize them accordingly which narrows down our domain search and foster the performance of the crawler.

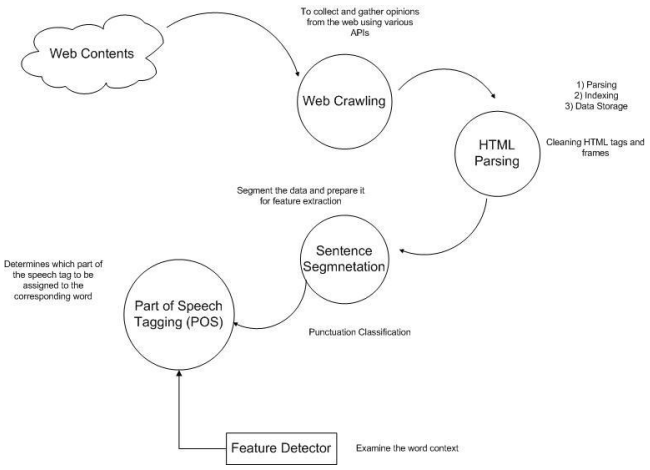

Fig. 1 Framework Stage 1

For the clarity of the framework demonstration, we break the HTML parsing out of the web crawling process, but technically all the parsing work happen inside our proposed focused web crawler.

A. Focused Web Crawler

Focused web crawler produce quality results and save more computing resources compare to the standard crawler. Our architecture depicts the way focused crawler picks the URL and start extracting it, and we show the looping techniques together with the dictionary mechanism that adds on the ability of being focused crawler.

Fig. 2 below shows the overall flowchart for our web crawler architecture. The design of the web crawler depends exclusively on the used programming language as there are many libraries that simplify the implementation details. For example, Java, Perl, Python and C# provide application programming interfaces and libraries for opening URL stream, buffer reading, queue construction, input/output operations and pattern recognition. Thus, implementation tools play a significant role in designing and developing a customized focused web crawler.

The first step in our implementation is to feed the crawler with URL input, and then we open an input stream based on that URL which is saved and added to the list by constructing a buffer to read the fed stream line by line. Append all the stream output to one string buffer to simplify the process of

text manipulation and control over the attached strings using loops and control statements. All the URLs are added to the linked list after testing certain conditions and constructing another linked list to save the visited URLs to avoid redundancy, hence it optimizes the retrieving process while fetching the URLs for page download.

In the following, we explain the implementation details shown in Fig. 2:

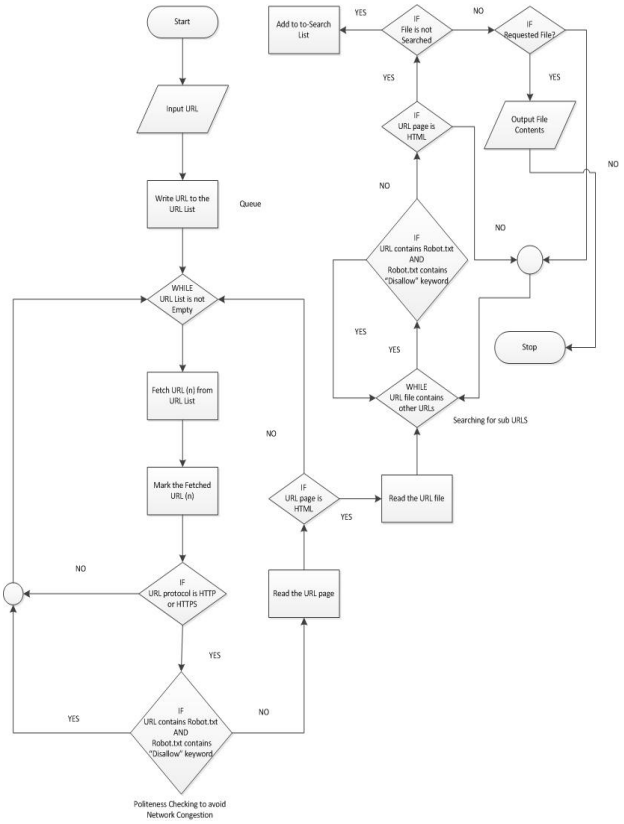

Fig. 2 Web Crawler Flowchart

- The method Fetch( ) is implemented with one parameters called links which provides the method with a list of URL's to start with
- Crawler will start visiting the first URL and then extract the remaining list of URLs and queue them in (ToVisit) list
- As long as the contents of URLs are eligible for politeness, protocol and type rules, the crawler will remain in the loop of extraction
- To be more polite, we added the (delay) parameter to ensure that the crawler waits for the specified amount of seconds before revisiting the same URL, the more seconds the less the traffic you put on the server and vice versa
- We provide additional control to the frontier manager to ensure that the crawler does not surf the entire web in an endless loop. Depth() is a method to maintain the number of visited links within the same domain. Breadth () is the other method that takes care of the other domains in the frontier manager
- The Breadth() method is capable of queuing the entire cache of web, but that will flood the memory and cause "Memory out of Buffer" error. That is where the priority check of the frontier comes in, whereby best first search is performed to pop the sub URLs in the frontier manager queue that has a limit of customized number.

B. Sentence Segmentation

Next in Fig. 1 is Sentence segmentation which breaks the sentences into smaller parts. Segmenting the text is dependent on the found delimiter such as, period, exclamation and question marks. Using regular expressions in the case of segmentation is effective enough to break the sentences into smaller chunks. Although the main difficulty in building sentence segmenter is the usage of punctuation and especially the period, in which it may refer to the sentence end or an abbreviation. Therefore, we segment using regular expressions and SRX for segmentation rules exchange to not confuse the nature of the punctuation usage. Immediately after the segmentation, a Bag of Words process known as tokenization is utilized to further subdivide the words as shown in Fig. 3 and prepare it for tagging process.

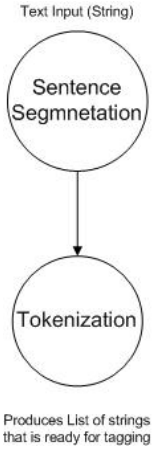

Fig. 3 Sentence Segmentation and Tokenization

C. Part of Speech Tagging (POS):

After the tokenization, words are labeled with tags according to their syntactic and grammatical structure as either noun, adjective, verb, adverb, pronoun, modal, etc. We use supervised learning tagging in which the POS tagger is pre-existed with all the words labeled based on a corpus just like the case in Brown corpus. We choose the Brown corpus as it is the oldest corpus with roughly one million words and concise tag set, another reason for choosing it is the easy accessibility to this corpus in various programming languages especially python and java.

The next stage as shown in Fig. 4 is all about the final word cleaning for classification. Stemming includes two separate functions; the first function is to reduce or derive the word back to its base form or root. For example, the words "typed"

and "typing" are converted back to the verb "type".

Second function is to remove or eliminate all stop words like 'the', 'that', 'this', 'to' and so forth as well as extras such as, apostrophe, commas and parentheses. According to the Wikipedia, there are several types of stemming algorithms including: Brute-force, the production technique, suffix-stripping, lemmatization, stochastic and hybrid approaches. We choose brute-force stemming algorithm as it can be built on the Brown corpus. The stemmer has to be connected to the tag set using a specific API to allow an easy access for chunking.

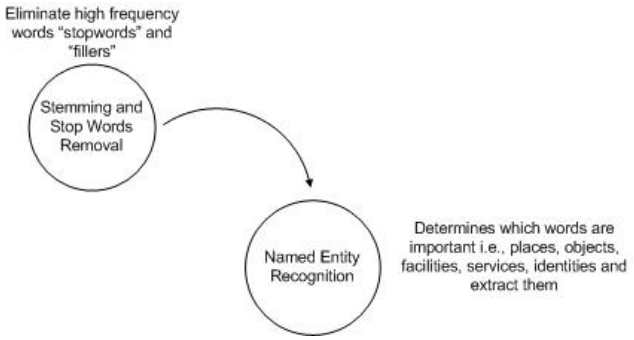

Fig. 4 Framework Stage 2

D.Text Pre-Processing and Processing:

The unstructured text, opinion or comment from the internet, is considered a raw text that needs few processing stages to produce meaningful data for polarity classification. After stage one in our framework, the text was already tagged based on the word individual and contextual meaning. However, in stage 2 the tagged words are further processed to extract the important entities for the typical evaluation.

TABLE I
SENTENCE SEGMENTATION AND POS TAGGING

I	LOVE	THE	Computer	Science
PNP	VVB	AT0	NN1	SENT
NP	-	NP		

With Table I we start explaining the underneath mechanism of segmentation and tagging. The first and second rows show the word level tokenization and POS tagging, while the third row demonstrates phrase level detection. For example, "The computer Science" phrase is detected as a Noun Phrase (NP) using a combination of regular expression rules with POS tagger where the rule is stated as:

```
  if ((optional determiner is detected & is
equal to (DT) tag) &
  ((count) number of adjectives are detected &
is equal to (JJ) tag) & (noun is detected & is
equal to (NN) tag)) = Identify the phrase as
NP
```

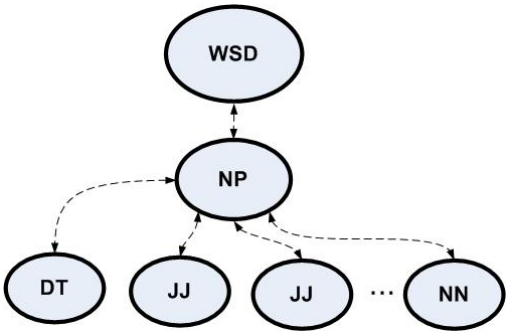

Fig. 5 WSD Tree

Fig. 5 depicts the significance of combining regular expression with POS tagging to produce contextual comprehension (WSD), whereby the word is interpreted based on the its independent and contextual meaning. After detecting all these features and built chunks of trees to examine the word context, we must prepare the processed text for NER by formatting the chunks into Inside-Outside-Begin (IOB) form.

In this scheme of format, every token is tagged with one of the three IOBs, the starting word in the chunk is (Begin), word in the middle is (Inside) and the ending word is (Outside). For instance, "I love the computer science" would be formatted as illustrated in Table II whereby the word inside a particular chunk is separated in terms of tagging from the word outside, therefore we consider the third column in Table II is the boundary between the two beginnings.

TABLE II
FORMATTING USING IOB

I	LOVE	THE	Computer	Science
PNP	VVB	AT0	NN1	SENT
NP	-	NP		
Begin (B)	Outside (O)	Begin (B)	Inside (I)	Inside (I)

E. Evaluating our Text Processing:

With the addition of IOB formatting to the line of processing together with parsing, sentence segmentation, tokenization, POS tagging and feature detection improved the performance and yield successful results as shown in Table III. Our method is a combination of various techniques entitled in as feature determiner or detector, while other text processing utilizes either classifier based or n-gram methodology. Comparison includes IOB formatting standard as it is embedded in all common implementations:

F. Named Entity Recognition (NER):

In this section we finish the discussion on the first and second stages by defining entities that are important for latter polarity classification. NER is necessary in our opinion mining framework to recognize the various entity types, hence simplifies the work of SentiWordNet to match the word or phrase to its particular category and find the polarity score easier.

Most of the NER methods are error prone when applied on huge amount of unstructured text with contents from various topics that might contain special terms corresponding to a

particular field. We are comparing the Malaysian universities based on the comments, feedbacks, posts and tweets collected from the online resources, therefore our domain is education with relative terms such as, subjects, campus facilities, courses, lecturers, tutors, labs, hostel and accommodation, sporting and lifestyle. Thus, we construct small corpus that serves our need to recognize the education domain words relative to its context and sentence level. Table IV shows a simple example for our types of named entity (corpus):

TABLE III
COMPARING TEXT PROCESSING RESULTS

Text Processing	Precision	Recall	F-Measure	IOB
1-gram	79.9%	86.8%	83.2%	92.9%
2-gram	82.3%	86.8%	84.5%	93.3%
Classifier	79.9%	86.7%	83.2%	92.9%
Feature	88.3%	90.7%	89.5%	95.9%

TABLE IV
LOOKUP TABLE FOR NAMED ENTITY

Key ID	Hash	Term	Type
u	ffab829a9a2352581ae530424ab83a08	University of malaya	Organization (university)
u	61a8c934d53c5a5bd2755930015b018b	multimedia university	Organization (university)
c	85285ed6d3ca6746428e22dfbb8571a9	computer science	Major (courses)
d	d1759cf9e6d6c674e93edab9aa8e124d	discrete mathematics	Subject (course)
U	66b67cf48eb78f0e0ae9902bc70d9e9a	urban	Lifestyle (sporting and lifestyle)

Through Table IV, we are able to recognize and extract mentioned words and phrases that are specific to our domain. This process makes the classification with SentiWordNet more accurate since it helps in recognizing the word considering its context and not only its independent meaning by using the relation formula and regular expression technique:

(X,α,Y) Where X and Y
=lookup table entity while α is the representing relation

G.Polarity Classification:

Fig. 6 shows the third phase of our framework. After processing the raw words and converting the unstructured text to a table format that can be utilized for polarity classification, the processed words are assigned a score value to express its positivity, negativity or objectivity (neutral). We are using the lexical resource SentiWordNet to find the corresponding polarity for the individual words. Baccianella et al. [2] explained that quantitative analysis and semi-supervised methods are the basis for SentiWordNet.

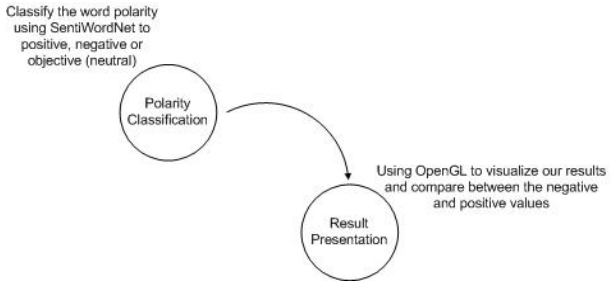

Fig. 6 Framework Stage 3

We chose the SentiWordNet as it is the largest lexical resource and freely available for research purposes. It fixes the word polarity in the lexicon by processing the word through eight ternary classifiers each with different classification behavior. The extracted entities of our domain are fed manually to SentiWordNet so that special terms are assigned accurate polarity based on the NER operation. The mechanism of feeding those manually extracted entities is by calculating the similarity between the fed entity and the WordNet 3.0 using Lee Raymond Dice:

$$s = \frac{2\,|\,X \cap Y\,|}{|\,X\,| + |\,Y\,|}$$

where (s) is the similarity variable measured by the strength of (X and Y)

In Fig. 6, the entire opinion mining framework in the education domain is finalized by presenting our findings on the Malaysian universities. The result presentation phase compares among the education institution based on the classified opinions earlier, whereby charts are displayed to easily visualize the comparison findings. We utilize the OpenGL libraries for visualization purpose due to its interactive outputs as shown in Fig. 7.

Next in Fig. 8 we combine all phases together and make up the complete framework for mining opinions from the internet in the education domain. This framework consisted of typical pre-processing stages including crawling, HTML parsing, sentence segmentation and POS tagging, unstructured text processing stages using stemming and NER, and last phase of polarity classification. The online resources of opinions and feedbacks can be extended dependent on the available bandwidth using simple APIs to get the latest information regarding the Malaysian universities.

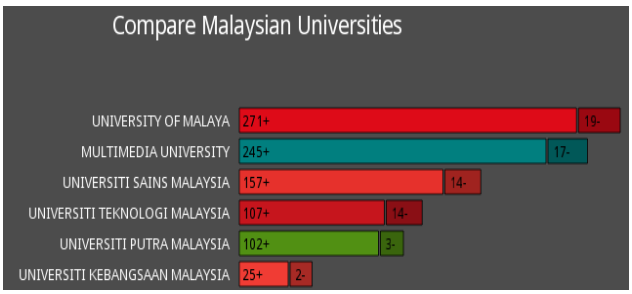

Fig. 7 OpenGL Output

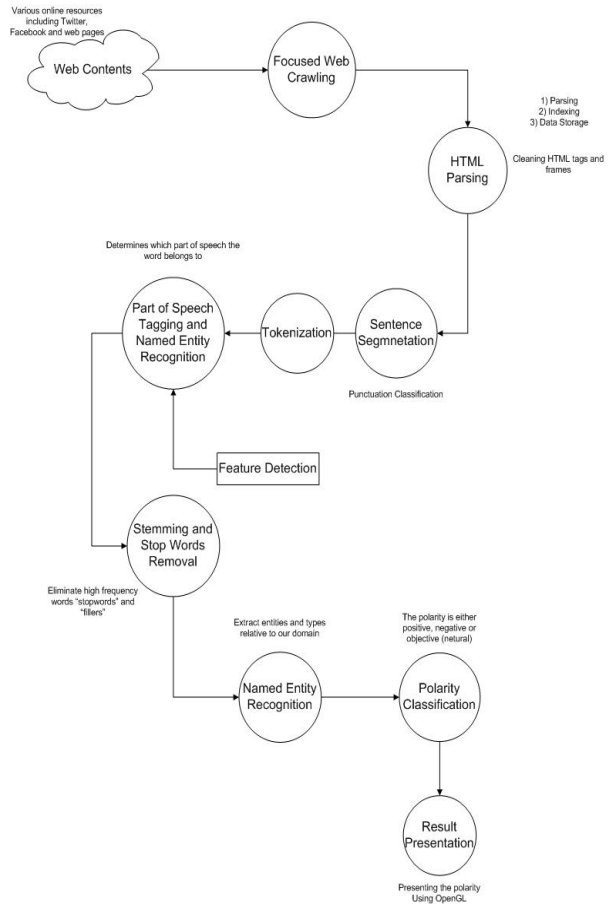

Fig. 8 Overall Framework

IV. Conclusion

We proposed opinion mining framework in the education domain to compare among the Malaysian universities based on the public chosen criteria. The framework is comprehensive and consists of three main stages of extraction, unstructured text processing and polarity classification. We collect the opinions using focused web crawler from the online platform specifically, twitter, Facebook and particular online web pages including blogs. Our framework is flexible enough to be embedded in future works of opinion mining to establish reliable benchmarks for chosen entities in a corresponding domain.

Acknowledgment

This work was supported by Project "Mining Opinions Using Combination of Word Sense and Bag of words approach for Educational Environments", funded by the Fundamental Research Grant Scheme of Malaysia.

References

[1] N. Archak, A. Ghose, and P.G Ipeirotis, "Deriving the Pricing Power of Product Features by Mining Consumer Reviews." *Management Science,* Vol 57(8), 2011, pp. 1485–1509.

[2] S. Baccianella, A. Esuli, and F. Sebastiani, "SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010.

[3] S. Park, M. Ko, J. Kim, Y. Liu, and J. Song, "The Politics of Comments: Predicting Political Orientation." *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW 2011)*, 2011, pp.113-122.

[4] M. Thelwall, D. Wilkinson, and S. Uppal. "Data Mining Emotion in Social Network Communication: Gender differences in MySpace." *Journal of the American Society for Information Science and Technology, Vol 61*, 2010: 190–199.