Using NLP Approach for Opinion Types Classifier

Mahmoud Othman[1*], Hesham Hassan[2], Ramadan Moawad[1], Amira M. Idrees[3]

[1] Department of Computer Science, Faculty of Computers and Information Technology, Future University, Cairo, Egypt.
[2] Department of Computer Science, Faculty of Computers and Information, Cairo University, Cairo, Egypt.
[3] Department of Information Systems, Faculty of Computers and Information, Fayoum University, Fayoum, Egypt.

* Corresponding Author. Tel.: 02-01113900394; email: ami04@fayoum.edu.eg

Manuscript submitted August 5, 2015; accepted November 20, 2015.
doi: 10.17706/jcp.11.5.400-410

Abstract: Information that are represented as text are either facts or opinions, whenever we need to make a decision, we often seek out the opinions of others which is one of the most influencing factors for our decisions. Traditionally, individuals can get opinions from friends and family while organizations use surveys, focus groups, opinion polls and consultants. Nowadays, opinions expressed through user generated content are considered as one of the important types of information which is available on the web, therefore, many resources have been emerged for expressing opinions including social media and others. This situation has revealed the necessity for robust, flexible Information Extraction (IE) systems, these systems have the availability to transform the web pages into program-friendly structures such as a relational database to reveal these opinions.

In this paper, we propose an approach to classify the opinions of a document or a set of documents considering an object. The approach has been implemented and applied on a dataset of opinions. The proposed system discover the opinions provided for an object in a document or set of documents. The system discovers different types of opinionated statements, including the opinionated, comparative, superlative, and non- opinionated. The system has been applied on a set of 4000 sentences, and the results has been evaluated using the standard metrics, they are True positive, True negative, False positive, False negative, Precision, Recall, and F-score. We also provided a comparison of the presented work with previous work that has been presented in the same field.

Key words: Opinion mining, opinion discovery, sentimental analysis, natural language processing.

1. Introduction

The internet presents a huge amount of useful information; this information is usually formatted for its users, which makes it difficult to extract relevant data from various sources. Whenever we need to make a decision, we often seek out the opinions of others; therefore, a need of information systems to seek for these opinions has become one of the vital targets in the current research fields.

Opinion mining [1], also called sentiment analysis, involves building a system to collect and examine opinions about the product or topic made in blog posts, comments, reviews or tweets. Several business intelligence applications across several domains are considering opinion mining as the key enabling technology for their business. Automated opinion mining often uses machine learning as well as natural language approaches.

There are several levels of sentimental analysis such as [2], [3]; Document-level which identifies if the document (e.g. product reviews, blogs, and forum posts) expresses an opinion, whether the opinion is positive, negative, or neutral; Sentence-level which identifies if a sentence is opinionated and whether the opinion is positive, negative, or neutral; and Attribute-level which extract the object attributes (e.g. image quality, zoom size) that are the subject of an opinion and the opinion orientations..

2. Background and Related Works

There are many approaches for opinion mining; lexicon-based approach and machine learning which are used to express sentiments and opinions. However, opinion mining is lacking of the resources especially multi-language resources.

In general, sentiment analysis is concerned with analyzing direction based text, determining whether a text is objective or subjective and whether a subjective text contains positive or negative sentiments is a common two-class problem that involves classifying sentiments as positive or negative. Additional variations include classifying sentiments as opinionated/subjective or factual/ objective. Some studies have attempted to classify emotions (such as happiness, sadness, anger, or horror) instead of sentiments.

Machine Learning Approaches The machine-learning approach [4]-[6], treats the sentiment-classification problem as a topic-based text classification problem. Any text classification algorithm can be employed, such as Naïve Bayes or support vector machines (SVMs) [7].

Lexicon Based Approach the Lexicon based approach [5], [8], performs classification based on positive and negative sentiment words and phrases contained. There are two types of techniques have been used in previous semantic orientation approach based sentiment classification research: corpus-based and dictionary-based.

As opinion discovery process, Murthy Ganapathibhotla and Bing Liu [9] proposed a research presenting the primary factors of the preferred entity in a comparative sentence which are the features being compared with the comparative words, which is extracted from the context of the opinions (or preferred entities). Popescu and Etzioni [6] worked on the explicit features in nous phrases, while Sobkowicz and *et al.* [9], used a corpus based technique which is applied on the politics domain, they used tweets as a dataset. Also Fei and *et al*. [10] used dictionary based as a technique. Finally some of previous researches merged between the two approaches, such as Yan Dang and *et al*. [11], who used SVM as a technique and merged it with a lexicon based approach, they used SVM as the classifier and for each test, they randomly choose 90 percent of the reviews as training data and the remaining 10 percent as testing data.

Our research is focusing on the detection of opinionated sentences and classify them and not considering the domain, specific entity or a specific feature (generality), some of related researches focus on a specific domain such as Sobkowicz and *et al*. [8], and others focused on the a specific type of opinion or sentence on specific entity such as, Murthy Ganapathibhotla and Bing Liu [9], also Popescu and Etzioni [10] worked on the explicit features in nous phrases only, O'Hara and *et al*. [12], has established a positive statistically significant correlation with the presence of adjectives. Thus the presence of adjectives is useful for predicting whether a sentence is subjective, i.e., expressing an opinion and they does not classify the type of the opinion whether it comparative or superlative only.

In our research project we trying to build a configurable framework which could applied on any domain, this paper is the first step by detecting the opinionated sentences and classifying them to (general, comparative or superlative) opinion using NLP approach.

3. Proposed System Framework

The proposed system aims to discover the opinion of the sentences in a document or a set of documents.

The system discovers different types of opinionated statements, including the Opinionated, comparative, superlative, and non- Opinionated.

The main approach of the system is to process the input document(s) targeting to tag each word in the document. Then in each sentence, the opinionated words are highlighted with providing a weight the highlighted words. Finally, the system defines the opinionated type of each sentence or a group of sentences according to the sentiment value highlighted opinionated words. Fig. 1 presents the main components of the proposed system, and Section 4 discusses the detailed description of each component in the proposed framework.

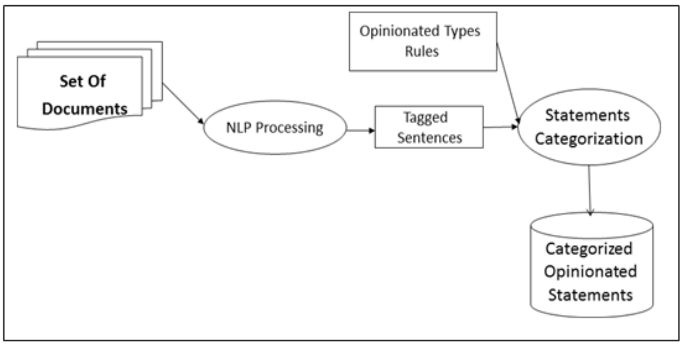


Fig. 1. General architecture of the proposed system.

4. Formal Description of the Proposed Approach

In this section we will describe the formal description of the proposed approach in details as follows:

4.1. Basic Definitions

If we consider d is a document, then we will define Document (d) as a set of all sentences.

$$\text{Document } (di) = \{s1, s2, s3 \ldots\ sn\}$$

Then, we will define Document (D) as a set of all documents d. The set of all documents Document (D) consists of all sentences in the documents.

$$\text{Document } (D) = \text{ Document } (d1) \cup \text{Document } (d2) \cup \text{Document } (d3) \ldots .. \cup \text{Document } (di)$$

The sentence consists of a set of tokens; then we will define sentence (s) as a set of all tokens w.

$$\text{Sentence } (sn) = \{w1, w2, w3, \ldots, wr\}$$

4.2. Formal Description for NLP Processing Phase

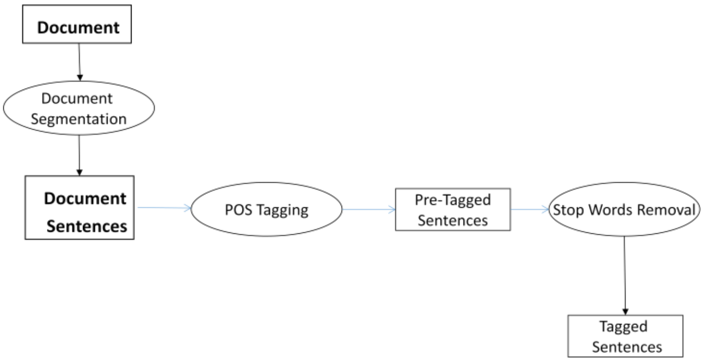


Fig. 2. General architecture of NLP processing phase.

The goal of NLP processing phase is to provide tags for each token in the document(s) with highlighting all stop words. Each word is then tagged with a suitable tag, these tags are applied using Stanford POS-Tagger [4]. The word with one of the tags that are described in Table 1 is considered a stop word.

NLP processing phase performs the target process in two steps, therefore this phase consists of two main components; they are: POS Tagger, and Stop Words Removal. Fig. 2 describes the architecture of NLP Processing phase, and the following subsections are the formal description of each component.

$$\text{Document } (di) = \{s1, s2, s3 \ldots\ sn\}$$

4.2.1. Formal description for document segmentation component

It is previously mentioned that the document is represented as a set of sentences. Therefore, in this component, the document is split into a group of sentences which are included in that document. This step is performed as a preparation step for the sentences in order to start the tagging procedure.

4.2.2. Formal description for POS tagger component

In any sentence, each token has its defined tag. The tokens in the sentence are tagged and the sentence is transformed into tagged sentences

$$\text{PSTagged } (sn\ ) = \{< w1/\ t1 >, < w2/\ t2 >, < w3/\ t3 >, <\ wr/\ tr > \mid < \frac{wn}{tn} > \quad are\ defined\ as\ a\ set\ of\ tokens\ with\ its\ attached\ tag\}$$

The previous definition means that we have a sequence of pairs; each pair consists of one token in the text and its tag value.

For example, if we have a sequence of tokens

Sn = “IPhone is better than blackberry”

We can define the set “PSTagged” by:

$$\text{PSTagged } (sn) = \{< \text{IPhone}/NN >, < \text{is}/VBZ >, < \text{better}/JJR >, < \text{than}/IN >, < \text{blackberry}/NN >\}$$

A set of a tagged document that contains all tagged sentences in a document is defined as follows:

The previous definition means that the set Dtagged(*di*) consists of the union of all sets of tagged tokens can be defined as follows:

Considering that the number of Documents is *i*, so a general formula which represents all tagged words in a document can be defined as follows:

$$D\text{Tagged}_{di} = \bigcup_{k=1}^{k=n} PS\text{Tagged}_{di}(S_k)$$

where:

*PS*Tagged (S_k) Mean the set of all tagged tokens in a sentence S_k in document *di*.

A general formula which represents all tagged words in all documents can be defined as follows

$$\boxed{D\text{Tagged } (D) \quad = \bigcup_{q=1}^{q=i} D\text{Tagged}_{dq}}$$

For example, a part of the document is as follows:

> “I bought an IPhone a few days ago it is such a nice phone the touch screen is really cool the voice quality is clear too it is much better than my old blackberry which was difficult to type with its tiny keys”

Then the output of this component is described in Fig. 3

I/PRP bought/VBD an/DT iPhone/NN a/DT few/JJ days/NNS ago/IN It/PRP is/VBZ such/JJ a/DT nice/JJ phone/NN The/DT touch/NN screen/NN is/VBZ really/RB cool/JJ The/DT voice/NN quality/NN is/VBZ clear/JJ too/RB It/PRP is/VBZ much/RB better/JJR than/IN my/PRP$ old/JJ Blackberry/NNP which/WDT was/VBD difficult/JJ to/TO type/NN with/IN its/PRP$ tiny/JJ keys/NNS

Fig. 3. An example for tagged text.

Table 1. Stop Word Tags

POS Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1, third
DT	Determiner	the
EX	existential there	there is
FW	foreign word	d'hoevre
IN	preposition/subordinating conjunction	in, of, like
LS	list marker	1)
MD	Modal	could, will
NN	noun, singular or mass	door
NNS	noun plural	doors
NNP	proper noun, singular	John
NNPS	proper noun, plural	Vikings
PDT	Pre-determiner	both the boys
POS	possessive ending	friend's
PRP	personal pronoun	I, he, it
PRP$	possessive pronoun	my, his
NNS	noun plural	doors
NNP	proper noun, singular	John
NNPS	proper noun, plural	Vikings
PDT	Pre-determiner	both the boys
POS	possessive ending	friend's
PRP	personal pronoun	I, he, it
PRP$	possessive pronoun	my, his
RB	Adverb	however
RP	Particle	give up
TO	To	to go
UH	Interjection	uhhuhhuhh
VB	verb, base form	take
VBD	verb, past tense	took
VBG	verb, gerund/present participle	taking
VBN	verb, past participle	taken
VBZ	verb, 3rd person sing. Present	takes
WDT	wh-determiner	which
WP	wh-pronoun	who, what
WP$	possessive wh-pronoun	whose
WRB	wh-abverb	where, when

4.2.3. Formal description for stop words removal component

In this component, all the stop words are removed from the document. The word is considered a stop word if it has a tag that is a member in the set of stop words.

Table 1 includes all tags that are members in the stop words set.

As each sentence contains all tagged words including the stop words, then a formula for the sentence in a document after removing the stop words can be described as follows:

$$S\text{Tagged}_{di}(S_n) = PS\text{Tagged}_{di}(S_n) - SW\text{Tagged}_{di}(S_n)$$

where:

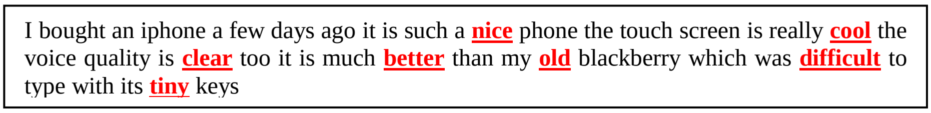
$S\text{Tagged}_{di}(S_n)$ Mean that the set of all tagged tokens in a sentence S_n in document *di* except the stop words

$SW\text{Tagged}_{di}(S_n)$ Mean that the set of all tagged tokens as stop words in a sentence S_n in document *di*.

The output of this component is a tagged document that contains all tagged words after removing the words that are tagged as stop words. Therefore, a general formula of the document after removing the stop words is as follows:

$$FD\text{Tagged}_{di} = \bigcup_{k=1}^{k=n} S\text{Tagged}_{di}(S_K)$$

Considering tagged text in Fig. 3, the output of the step will be the tagged text after high following the example in Fig. 3, the words that are not stop words are highlighted as described in Fig. 4.



I bought an iphone a few days ago it is such a **nice** phone the touch screen is really **cool** the voice quality is **clear** too it is much **better** than my **old** blackberry which was **difficult** to type with its **tiny** keys

Fig. 4. Stop words removal component output.

4.3. Statement Categorization Phase

The goal of Statement Categorization Phase is to categorize each statement to one of four sentimental categories; Table 2 includes these categories, a simple description, the part of speech defining each category, and an example for each type.

Table 2. Description of Sentence Categories

Sentimental Category	Description	POS
Non-Opinionated Statement	Any statement that doesn't state any opinion.	--
Comparative Opinionated Statement	Any statement that state opinion by comparing two or more objects.	JJR, RBR
Superlative Opinionated Statement	Any statement that state opinion by comparing more than one objects and considered the optimal.	JJS, RBS
Opinionated Statement	Any statement that state opinion on a certain object	JJ

Table 3. POS Tags for Tokens That Are not Stop Words

POS Tag	Description	Example
JJ	Adjective	Big
JJR	adjective, comparative	Bigger
JJS	adjective, superlative	Biggest
RBR	adverb, comparative	Better
RBS	adverb, superlative	Best

Each sentence type is defined by a set of POS tags; the tags for these types are described in Table 3.

A formal description of the set of tokens that are not stop words in a document is as follows:

$$STagged(di) = NSTagged\ (di)\ \cup SSTagged\ (di)\ \cup CSTagged\ (di) \cup OSTagged\ (di)$$

where:

*NS*Tagged(*di*) is the set of all tokens that are of type non-Opinionated.

*CS*Tagged(*di*) is the set of all tokens that are of type Comparative Opinionated.

*SS*Tagged(*di*) is the set of all tokens that are of type Superlative Opinionated.

*OS*Tagged(*di*) is the set of all tokens that are of type Opinionated.

The sentence is considered of type Non-Opinionated as default, then statement categorization phase determines the type of the sentence according to the algorithm described in Fig. 5.

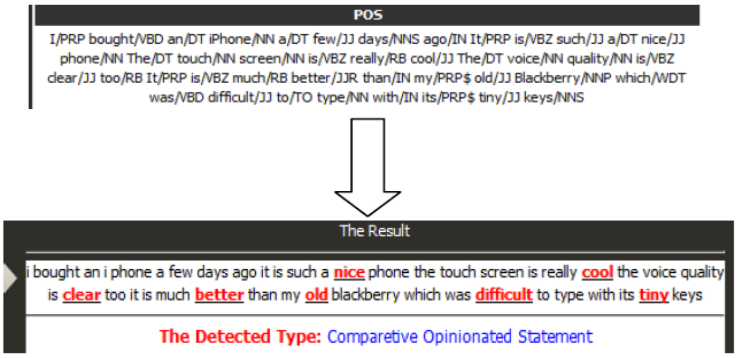
```
Set Stype to "Not Opinionated Statement";
If(txtPOS.contains("/JJS") OR    txtPOS.contains("/RBS"))
        Stype="Superlative Opinionated Statement";
else if(txtPOS.contains("/JJR") OR    txtPOS.contains("/RBR"))
        Stype="Comparative Opinionated Statement";
Else if(txtPOS.contains("/JJ"))
                Stype="Opinionated Statement";
```

Fig. 5. An example for tagged text.

4.4. System Output

To finalize the system framework description, we present four examples that produce different types of opinions. Fig. 6 presents the output for the previously mentioned example. The output consists of two parts, the first part is the sentence after tagging each token, and the second part is the sentence after considering only opinionated words and the type of the sentence.

In the example of Fig. 6, the system has inferred that the sentence is of type “comparative opinionated statement”.



POS

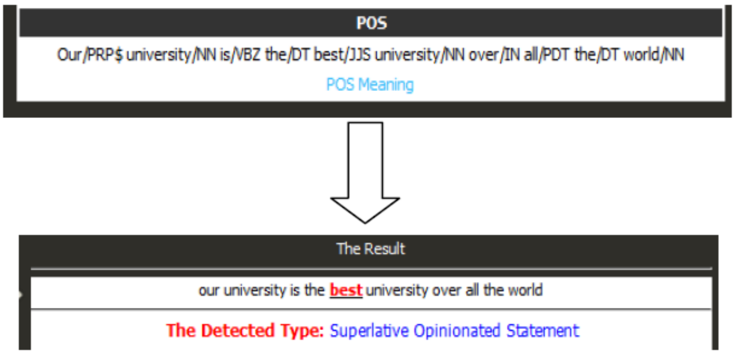
I/PRP bought/VBD an/DT iPhone/NN a/DT few/JJ days/NNS ago/IN It/PRP is/VBZ such/JJ a/DT nice/JJ phone/NN The/DT touch/NN screen/NN is/VBZ really/RB cool/JJ The/DT voice/NN quality/NN is/VBZ clear/JJ too/RB It/PRP is/VBZ much/RB better/JJR than/IN my/PRP$ old/JJ Blackberry/NNP which/WDT was/VBD difficult/JJ to/TO type/NN with/IN its/PRP$ tiny/JJ keys/NNS

The Result

i bought an i phone a few days ago it is such a **nice** phone the touch screen is really **cool** the voice quality is **clear** too it is much **better** than my **old** blackberry which was **difficult** to type with its **tiny** keys

The Detected Type: Comparetive Opinionated Statement

Fig. 6. Output of the proposed system.



POS

Our/PRP$ university/NN is/VBZ the/DT best/JJS university/NN over/IN all/PDT the/DT world/NN

POS Meaning

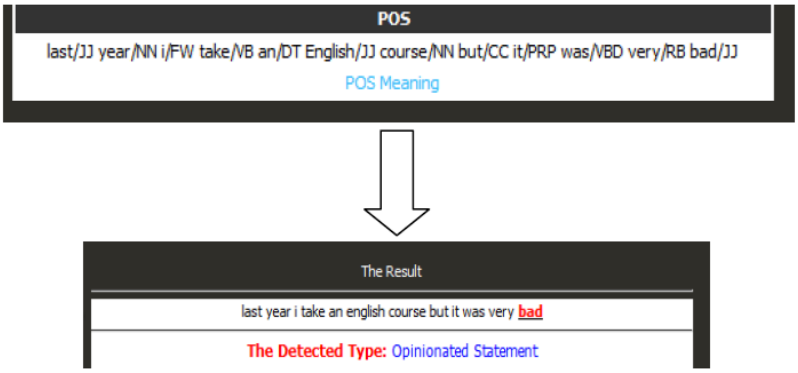
The Result

our university is the **best** university over all the world

The Detected Type: Superlative Opinionated Statement

Fig. 7. Second example for the proposed system.

More examples are presented in the Figs. 7, 8, and 9 presenting different types of sentences that are discovered from the proposed systems.

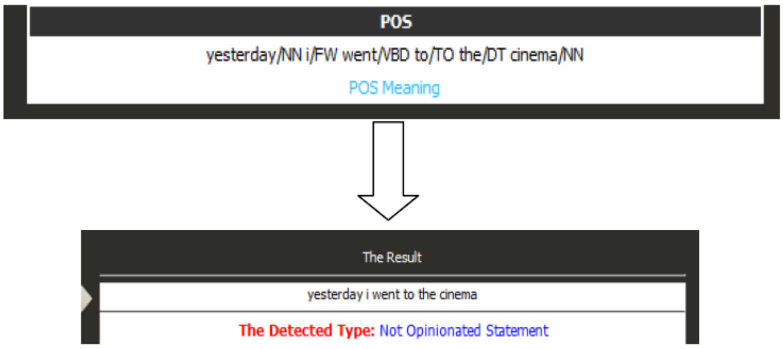


last/JJ year/NN i/FW take/VB an/DT English/JJ course/NN but/CC it/PRP was/VBD very/RB bad/JJ

last year i take an english course but it was very **bad**

The Detected Type: Opinionated Statement

Fig. 8. Third example for the proposed system.



yesterday/NN i/FW went/VBD to/TO the/DT cinema/NN

yesterday i went to the cinema

The Detected Type: Not Opinionated Statement

Fig. 9. Fourth example for the proposed system.

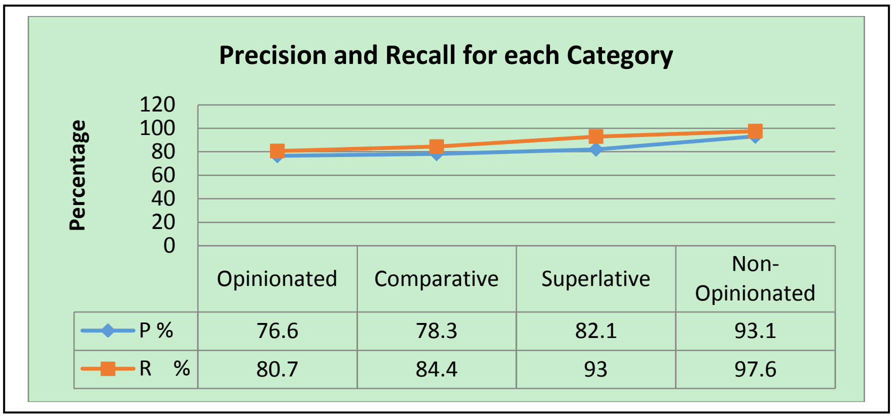
5. Evaluation and Discussion

As we introduced in the previous section, the proposed system aims to discover the opinion of the sentences in a document or a set of documents. The system discovers different types of opinionated statements, including the Opinionated, comparative, superlative, and non-Opinionated; we measured precision and recall for the classification over four different types of categories summarized in Table 4.

Table 4. Number of Categories & Statements That Is Used in the Evaluation

Category	Number of statements
Opinionated	1500
Comparative	1000
Superlative	1000
Non- opinionated	500
Total	4000 Statement

We used 4000 statements as test data which are used in [9] for determining context (aspect) dependent sentiment words, this test data are classified by four categories as mentioned above, for the first category, “opinionated” 920 are classified correctly as opinionated but 280 statements classified incorrectly as opinionated due to the irregularity of the statements. 30 opinionated statements were not classified as opinionated; also 220 statements were not classified as opinionated which is correct. According to the previous output analysis, the precision of the first category was 76.6% and the recall was 80.7%. We have performed all the previous process to all other three categories and the result summarized in figure 10, finally the average accuracy of the whole proposed system according to the data set used was 85.6%. but , M. Ganapathibhotla, B. Liu [9] use only the comparative statement using manual rules as the attribute for naïve Bayesian model and gave a precision of 82% and a recall of 81% (F-score = 81%) for identification of gradable comparative sentences. they also tried various other techniques, e.g., SVM but the results were all poorer due to space limitations.



	Opinionated	Comparative	Superlative	Non-Opinionated
◆ P %	76.6	78.3	82.1	93.1
■ R %	80.7	84.4	93	97.6

Fig. 10. Precision and recall.

Also Minqing Hu, *et al*. [13], conducted their experiments using the customer reviews of five electronics products: 2 digital cameras, 1 DVD player, 1 mp3 player, and 1 cellular phone. The reviews were collected from Amazon. Products in these sites have a large number of reviews. Each of the reviews includes a text review and a title. Additional information available but not used in this project includes date, time, author name and location (for Amazon reviews), and ratings. they uses adjectives as opinion words and this limit the opinion words extraction to those sentences that contain one or more product features because they only interested in customers' opinions on these product features with average precision 69.3 and average recall 64.2, In this paper we conducted our experiments on the same dataset in a single category MP3 Players Products. Our precision was 81.6 and recall 80.1 but was in [13] for MP3 players 65.3, 60.7 respectively.

6. Conclusion and Future Work

Several opinion Mining and sentimental analysis approaches have been developed to analyze comments, or tweets that are related to services and products. There are more than one approach used such as machine learning approaches and lexicon based approaches. In this work, we proposed a system the aims to discover the opinion of the sentences in a document or a set of documents. The system discovers different types of opinionated statements, including the Opinionated, comparative, superlative, and non- Opinionated. The system has been evaluated by applying it on a set of 4000 sentences and the results and the approach have been compared with other related work in the same field to prove its applicability and advancement.

Our future work is the extension of this work to include multi-language opinions, improve the performance to construct an opinion search engine in different domains and try to propose a configurable approach for opinion mining.

References

[1] Mikalai, T., & Themis, P. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery Journal, 24*, 478-514.

[2] Cruz, F., Troyano, J., Ortega, F., & Vallejom, C. (2013). Long autonomy or long delay? The importance of

domain in opinion mining. *Expert Systems with Applications, 40.*

[3] Lei, Z., & Bing, L. (2011). Sentiment analysis and opinion mining. *Introduction and Survey Book*, Morgan & Claypool.

[4] Toutanova, K., & Manning, C. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpor*, EMNLP/VLC.

[5] Othman, M., Hassan, H., Moawad, R., & El-Korany, A. (2014). Opinion mining and sentimental analysis approaches: A survey. *Life Science Journal*, *11(4),* 321-326.

[6] Ana-Maria, P., & Etzioni, O. (2005). Extracting product features and opinions from reviews. *Proceedings of the Conference on Empirical Methods in Natural Language Processing, (EMNLP2005)*.

[7] Jebaseel, A., & Kirubakaran, E. (2012). M-learning sentiment analysis with data mining techniques. *International Journal of Computer Science and Telecommunications, 3*.

[8] Sobkowicz, P., Kaschesky, M., & Bouchard, G. (2012). Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. *Government Information Quarterly*, *29*, Elsevier.

[9] Ganapathibhotla, M., & Liu, B. (2008). Mining opinions in comparative sentences. *Proceedings of the 22nd International Conference on Computational Linguistics,* Manchester.

[10] Fei, G., Liu, B., Hsu, M., Castellanos, M., & Ghosh, R. (2012). A dictionary-based approach to identifying aspects implied by adjectives for opinion mining. *Proceedings of COLING* (pp. 309–318).

[11] Dang, Y., Zhang, Y., & Chen, H. (2010). A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *IEEE Intelligent Systems*, *25.*

[12] O'Hara, T., Wiebe, J., & Bruce, R. (1999). Development and use of a gold standard data set for subjectivity classifications. *Proceeding of ACL*.

[13] Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 168-177).

Mahmoud Othman is Assistant Lecturer at the Faculty of Computers and Information Technology, Future University. He graduated from the Faculty of Computers and Information, Cairo University with grade excellent with majoring in computer science and a minor in information systems. He ranked from the top ten students in computer science department. He got a master degree from Cairo University after two years graduation. He received two awards as an outstanding TA for the academic years 2011-2012 and 2013-2014. Now he prepares his Ph.D. degree at the Faculty of Computers and Information, Cairo University.

Hesham Hassan is Professor in the Faculty of Computers and Information, Cairo University. He is the organizer of Software Engineering Special Program, Cairo University. His research interests include knowledge discovery, text mining, opinion mining, sentimental analysis, question answering, data warehousing, cloud computing, and other topics related to computer science.

Ramadan Moawad obtained his BSc from Military Technical College in Electric Engineering and M.Sc degree from Military Technical College in Computer Engineering. He obtained his PhD from ENSAE College, France in software engineering. He taught several courses in CS and CE IN several institutions including the American University in Cairo, the Military Technical College, Cairo University and the Arab Academy for Science and Technology. He joined Future University in 2011 and currently working as Vice-Dean of FCIT. He published over 50 papers in different journals and conferences locally and internationally. His research interest is software engineering and software quality assurance. He has refereed several papers in IEEE Transactions in Software Engineering Journal and the International Journal of Software Engineering (IJSE).

Amira M. Idrees is an Associate Professor in the Faculty of Computers and Information, Fayoum University.

Her research interests include knowledge discovery, text mining, opinion mining, sentimental analysis, cloud computing, software engineering, and data warehousing