Statistical Modeling for Opinion Mining: A Review

Deepti Gupta[1], Arivind Singh Chandel[2]
[1]O.P Jindal Institute, Raigarh(C.G)
[2]O.P Jindal Institute, Raigarh(C.G)
(E-mail: deeptigupta.it@gmail.com)

***Abstract*—**Opinion classification is the process of using NLP, statistics, or machine learning methods to extract, identify, or otherwise characterize the sentiment content of a text unit. Based on a sample of tweets, how are people responding to this ad campaign/product release/news item? There are several application of opinion mining such as on business intelligence, Politics/political science, Law/policy making, Sociology, Psychology etc. By use of digital platform administration can collect response from consumer and by means of applying opinion mining technique a useful information from user collected data. In this paper we have given a brief review over opinion mining and given tabular comparison among different opinion classification technique based on accuracy.

Keywords*—TPR,FNR,ML,NL,SVM ANN.*

I. INTRODUCTION

Extracting information from social networking post, news articles and other texts is a significant application task for natural language processing technology. In the previous couple of years, web related document are accepting awesome consideration as another medium that depicts singular encounters and conclusions, as symbolized by the new word, for example, Blog news coverage" or Consumer produced media (CGM)". This circumstance is generating much enthusiasm for innovations for consequently separating or breaking down sincere beliefs from web related document, for example, posts on message board and weblogs. Such advancements can be a contrasting option to customary poll based social or client inquire about and would likewise profit Web clients who look for surveys on certain customer results of their advantage.Opinion mining, also recognized as sentiment classification, is the study and application of text-related computational methods, such as natural language processing (NLP) and text mining, to identify people's opinions about products, movies, news, etc.

Most organizations have led reviews and statistical surveying to get customers' supposition about their items. The requirement for opinion examination develops considerably in the time of the web. With the quick development of web based business, the online buy of products is predominant, and the buyers regularly compose audits [Liu 2012]. Such online audits assume a vital part in others' basic leadership, particularly with dynamic web-based social networking. As of late, more individuals have investigated such audits and remarks on merchandise on the web before they purchase an item. Because of this marvel, neighborhood stores as well as undertakings need to keep up positive surveys of their items and administrations.

Automatic sentiment classification and examination methods are useful in numerous applications with psychosomatic basis. For instance, it can be magnificently applied to acquire user favorites and comforts from users' personal written text and vocalizations. These methods are frequently considered in the field of the domain of behavior or mood modeling and customer response analysis. Likewise, e-learning systems can assistance from affective tutoring approaches.

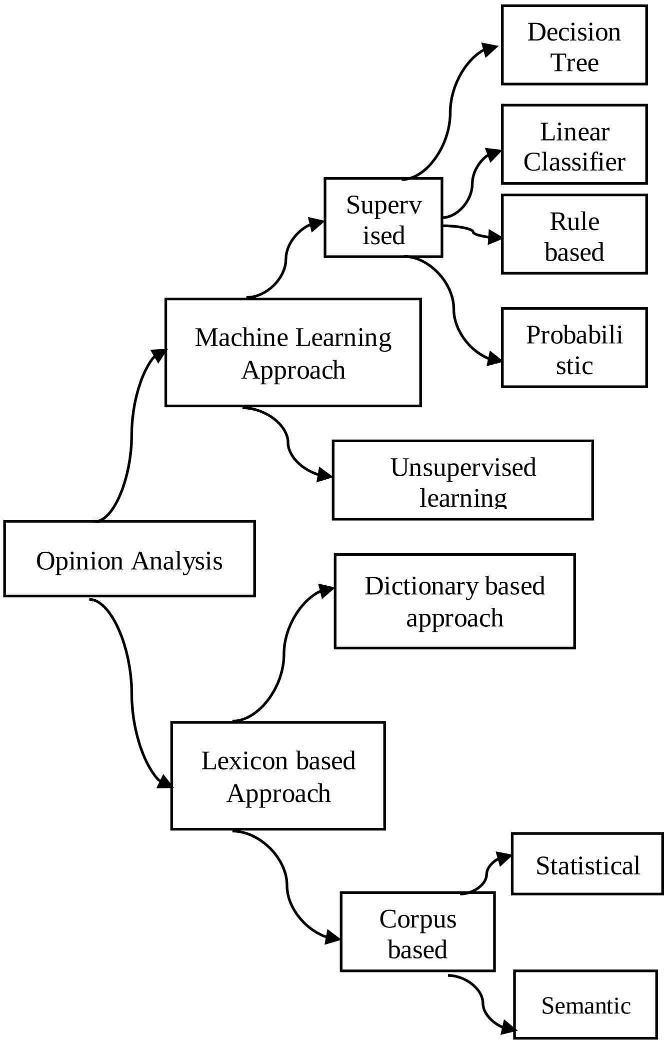

Fig.1- Opinion Classification Techniques

There are several fields where automatic sentiment classification is worthwhile for example there are several e-commerce websites in which they ask for feedback, product review etc. The user or customer can give diverse thoughts for different aspects of the identical entity like this sentence ‘‘the voice feature of this phone is not good, but the battery life is long’’.

These criticisms/review are significant to the business owners as they can take commercial resolutions or decisions agreeing to the scrutiny results of users' sentiments about their products. The reviews sources are primarily review sites. Sentiment analysis (SA) is not only efficient on merchandise reviews but can also be pragmatic on stock markets, news tutelages, or political discussions.

A. Applications

1) Based on a sample of tweets, how are people responding to this ad campaign/product release/news item?
2) How have bloggers' attitudes about the president changed since the election?
3) Identifying child-suitability of videos based on comments.
4) Identifying (in) appropriate content for ad placement.
5) Use SA to
   a) Search the web for opinions and reviews of this and competing laptops. Blogs, Opinions, amazon, tweets, etc.
   b) Create condensed versions or a digest of consensus points

Further in this paper in section-II we will go through different literature, in section-III we will give a tabular comparison among different literature, in section IV challenges involved in opinion classification, in section V we will describe general steps involved in how opinion can be analyzed from text, at last we will conclude our study.

II. LITERATURE SURVEY

B. Vamshi Krishna et. al. presented model which is based on feature extraction of the products and their degree of polarity is converted into fuzzy sets. The proposed model is used to analyze user opinions and reviews posted on social media websites and helps users in decision making to buy products and organizations to recommend products online. Model proposed in the paper utilizes machine learning techniques and fuzzy approach for opinion mining and classification of sentiment on textual reviews. The goal is to automate the process of mining attitudes, opinions and hidden emotions from text [Springer 2018].

Y. Wang et. al. said that online reviews are acknowledged as an important source of product information when customers make purchasing decisions. However, in the era of information overload, product review data on the Internet are too abundant and contain much irrelevant information. This makes it difficult for customers to find useful reviews. To solve this issue, some e-commerce websites provide keywords for product reviews, but these are generated beforehand and have the potential to distort customers' opinions of products. This paper presents an automatic keyword extraction method based on a bi-directional long short-memory (LSTM) recurrent neural network (RNN). The results of experiments conducted on product reviews obtain by data-crawling jd.com show that the proposed approach has a very high accuracy of keyword extraction. This can help reduce human annotation efforts in ecommerce. Author also concluded that Product review data provide valuable information for customers to better understand the product and make purchasing decision. However, information overload issue persists in product review data. This paper presents a LSTM RNN based approach to extract keywords from product review text. The results make customers quickly get the review opinion about the product and could greatly facilitate consumer decision making process. Provided with the set of keywords generated from product review texts, relevant research can be conducted [IEEE 2017].

Sandra Garcia Esparza et. al. said that Real-time web (RTW) services such as Twitter allow users to express their opinions and interests, often expressed in the form of short text messages providing abbreviated and highly personalized commentary in real-time. Although this RTW data is far from the structured data (movie ratings, product features, etc.) that is familiar to recommender systems research, it can contain useful consumer reviews on products, services and brands. This paper describes how Twitter-like short-form messages can be leveraged as a source of indexing and retrieval information for product recommendation. In particular, we describe how users and products can be represented from the terms used in their associated reviews. An evaluation performed on four different product datasets from the Blippr service shows the potential of this type of recommendation knowledge, and the experiments show that our proposed approach outperforms a more traditional collaborative-filtering based approach [Elsevier 2011].

Isa Maks et. al. presents a lexicon model for the description of verbs, nouns and adjectives to be used in applications like sentiment analysis and opinion mining. The model aims to describe the detailed subjectivity relations that exist between the actors in a sentence expressing separate attitudes for each actor. Subjectivity relations that exist between the different actors are labeled with information concerning both the identity of the attitude holder and the orientation (positive vs. negative) of the attitude. The model includes a categorization into semantic categories relevant to opinion mining and sentiment analysis and provides means for the identification of the attitude holder and the polarity of the attitude and for the description of the emotions and sentiments of the different actors involved in the text. Special attention is paid to the role of the speaker/writer of the text whose perspective is expressed and whose views on what is happening are conveyed in the text. Finally, validation is provided by an annotation study that shows that these subtle subjectivity relations are reliably identifiable by human annotators [Elsevier 2012].

Giulio Angiani et. al. said that Sentiment Analysis has become one of the most interesting topics in AI research due to its promising commercial benefits. An important step in a Sentiment Analysis system for text mining is the preprocessing phase, but it is often underestimated and not extensively covered in literature. In this work, our aim is to highlight the importance of preprocessing techniques and show how they can improve system accuracy. In particular, some different preprocessing methods are presented and the accuracy of each of them is compared with the others. The purpose of this comparison is to evaluate which techniques are effective. In this paper, we also present the reasons why the accuracy improves, by means of a precise analysis of each method. Author also concluded that Text preprocessing is an important phase in all relevant applications of data mining. In Sentiment Analysis, in particular, it is cited in virtually all available research works. However, few works have been specifically dedicated to understanding the role of each one of the basic preprocessing techniques, which are often applied to textual data [Springer 2016].

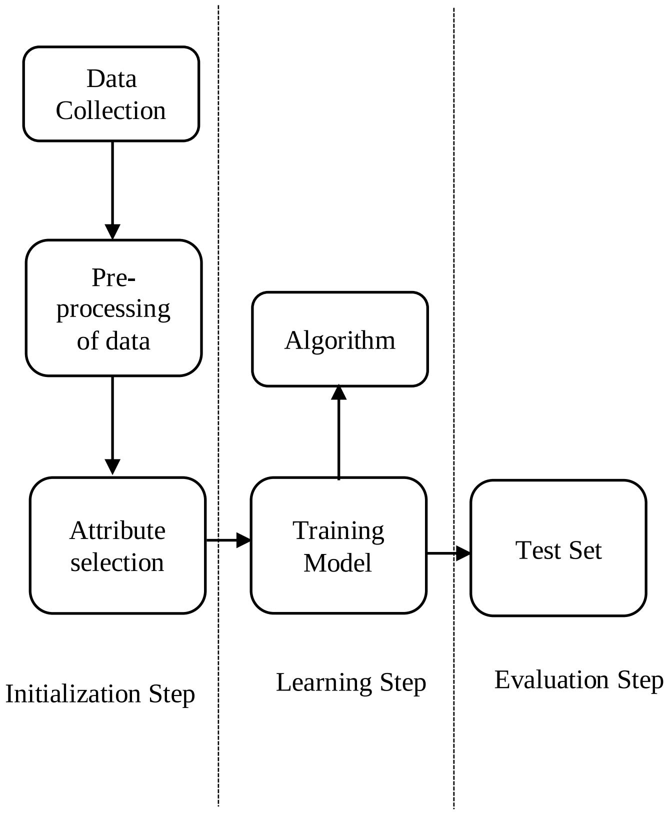

Fig.2- Steps in SA [Giulio Angiani et. al. Springer 2016]

Erik cambria et.al. said that Next-generation sentiment-mining systems need broader and deeper common and commonsense knowledge bases, together with more brain inspired and psychologically motivated reasoning methods, to better understand natural language opinions and, hence, more efficiently bridge the gap between (unstructured) multimodal information and (structured) machine-process able data[IEEE 2016].

III. COMPARISON

The comparison of work done by different authors are presented in table III

TABLE I.Comparison OF work done by different authors

S. No.	Author/Title/Publication/Year	Algorithm used	Description
1.	B. Vamshi Krishna, Ajeet Kumar Pandey and A. P. Siva Kumar/Feature Based Opinion Mining and Sentiment Analysis Using Fuzzy Logic/Springer 2018	Machine learning techniques and fuzzy approach	The proposed model is used to analyze user opinions and reviews posted on social media websites and helps users in decision making to buy products and organizations to recommend products online. **F-Score:- 0.36**
2.	Y. Wang, J. Zhang/Keyword Extraction from Online Product Reviews Based on Bi-Directional LSTM Recurrent Neural Network/IEEE 2017	Bi-Directional LSTM	Paper presents an automatic keyword extraction method based on a bi-directional long short-memory (LSTM) recurrent neural network (RNN). **Accuracy:- 93%**
3.	Sandra Garcia Esparza , Michael P. O'Mahony, Barry Smyth/Mining the real-time web: A novel approach to product recommendation/Elsevier 2011	Real time web mining	Paper investigates how user-generated micro-blogging messages can be used as a new source of recommendation knowledge. We have proposed an approach to represent users and products based on the terms in their associated reviews using techniques from the information retrieval community. **Precision:- 84%**
4.	Isa Maks , Piek Vossen/A lexicon model for deep sentiment analysis and opinion mining applications/Elsevier 2012	Lexicon model	Paper presents a lexicon model for the description of verbs, nouns and adjectives to be used in applications like sentiment analysis and opinion mining. The model aims to describe the detailed subjectivity relations that exist between the actors in a sentence expressing separate attitudes for each actor.
5.	Rodrigo Moraes, João Francisco Valiati, Wilson P. Gavião Neto/Document-level sentiment classification: An empirical comparison between SVM and ANN/Elsevier 2013	SVM and ANN	Author experiments indicated that ANN produce superior or at least comparable results to SVM's. Especially on the benchmark dataset of Movies reviews, ANN outperformed SVM by a statistically significant difference, even on the context of unbalanced data.

6.	Tarik S. Zakzouk/Comparing text classifiers for sports news/Elsevier 2012	-	In this paper, author revisit this field using both commodity software and hardware to show progress of both efficiency and effectiveness of a group of ML-based methods in classifying Cricket sports news articles.
7.	Ngoc Phuong Chau, Viet Anh Phan, Minh Le Nguyen/Deep Learning and Sub-Tee Mining for Document Level Sentiment Classification/KSE 2016	LSTM + GRNN model	The association between all words in a sentence and all sentences in a document is captured by LSTM and GRNN, respectively. Document sentiment classification experiment is conducted on multi-domain sentiment dataset. The elimination of outliers leads to higher performance in this model• In experiment, the proposed method achieves improvements in term of accuracy in a range of **0.14% - 6.93%** over LSTM + GRNN model.
8.	Ivo Danihelka et. al./Associative Long Short-Term Memory/arXiv 2016	Associative LSTM	System in contrast creates redundant copies of stored information, which enables retrieval with reduced noise. Experiments demonstrate faster learning on multiple memorization tasks.

IV. CHALLENGES IN OPINION CLASSIFICATION

People are subjective creatures and assessments are imperative. Having the capacity to interface with individuals on that level has numerous points of interest for data frameworks. There are some challenges in opinion classification:

- People express opinions in complex ways
- In opinion texts, lexical content alone can be misleading
- Intra-textual and sub-sentential reversals, negation, topic change common
- Rhetorical devices/modes such as sarcasm, irony, implication, etc

V. GENERAL STEPS FOR OPINION CLASSIFICATION

A. Preprocessing:

1) Cleaning operation: In this phase consists removing unimportant or disturbing elements for the next phases of analysis and in the normalization of some misspelled words. Remove the vowels repeated in sequence at least three times, because by doing so the words are normalized: for example, two words written in a different way (i.e. cooooool and cool) will become equals.
2) Emoticon: This phase reduces the number of emoticons to only two categories: smile positive and smile negative as given:

TABLE II. LIST OF EMOTICONS

Smile positive	Smile negative
0:-)	>:(
:)	;(
:D	>:)
:*	D:<
:o	:(
:P	:j
;)	>:/

3) Dictionary: This phase countenances us to substitute slang with its formal meaning (i.e., *l8* with late), using dictionary constructed. This is very important to reduce the noise in text and improve the overall classification performances.
4) Stemming: Stemming techniques put word variations like \great", \greatly", \greatest", and \greater" all into one bucket, effectively decreasing entropy and increasing the relevance of the concept of \great". In other words, Stemming allows us to consider in the same way nouns, verbs and adverbs that have the same radix.
5) Spam Words: Stop words are words which are filtered out in the preprocessing step. These words are, for example, pronouns, articles, etc. It is important to avoid having these words within the classifier model, because they can lead to a less accurate classification.

B. Feature Extraction:

N-gram: An n-gram model is a type of probabilistic language model for predicting the next word conditioned on a sequence of previous words using Markov models. N-gram of size 1 is referred to as unigram, size 2 as bigram, and size 3 as trigram. Since n-grams are used for capturing dependencies between single words that stay in a text sequentially.

EXAMPLE:

Text: “Honesty is the best policy.”

Unigrams: “honesty”, “is”, “the”, “best”, “policy”.

Bigrams: “honesty is”, “is the”, “the best”, “best policy”.

Trigrams: “honesty is the”, “is the best”, “the best policy”.

C. Classification:

There are different classification algorithms which depicted in fig.-1.

1) STEP1- Finding “Bag of Words” from training set” (I, loved, the, movie, hated, a, great, poor, acting, good)

TABLE III. FINDING BAG OF WORDS

Document ID	Text	Class
1	I loved the movie	EXT
2	I hated the movie	
3	A great movie. Good movie. Poor acting	AGR
4	poor acting	EXT
5	Great acting, good movie	EXT
6	Great acting	

2) STEP2- Reducing dimension of input set: remove rows which having missing values.

TABLE IV. REDUCING DIMENSION OF TEXT

Document ID	Text	Class
1	I loved the movie	EXT
3	A great movie. Good movie. Poor acting	AGR
4	poor acting	EXT
5	Great acting, good movie	EXT

3) STEP3-Convert document into feature

TABLE V. SELECTING FEATURE OF TEXT DOCUMENT

Document ID	I	Loved	The	Movie	Hated	A	Great	Poor	Acting	good	Class
1	1	1	1	1							EXT
3	1			2		1	1	1	1	1	AGR
4								1	1		EXT
5							1		1	1	EXT

VI. CONCLUSION

In this paper, the task of opinion mining has been done as per sentiment present in text. The decomposition of problem of opinion mining into the following series of subtasks has been done for:

- Extraction of opinions in a structured form
- Determination of semantic orientation i.e. to each extracted opinion positive, negative, or neutral semantic orientation is assigned.

At last we can conclude that opinion mining involve different steps as preprocessing, feature extraction and classification, accuracy of algorithm depends on algorithm accuracy which are used for above steps. Hence, by applying better technique we can increase the accuracy and we saw ANN performed well.

REFERENCES

[1] B. Vamshi Krishna, Ajeet Kumar Pandey and A. P. Siva Kumar Feature Based Opinion Mining and Sentiment Analysis Using Fuzzy Logic Springer 2018

[2] Y. Wang, J. Zhang Keyword Extraction from Online Product Reviews Based on Bi-Directional LSTM Recurrent Neural Network IEEE 2017

[3]Sandra Garcia Esparza , Michael P. O'Mahony, Barry Smyth Mining the real-time web: A novel approach to product recommendation Elsevier 2011

[4]Isa Maks , Piek Vossen A lexicon model for deep sentiment analysis and opinion mining applications Elsevier 2012

[5]Rodrigo Moraes, João Francisco Valiati, Wilson P. Gavião Neto Document-level sentiment classification: An empirical comparison between SVM and ANN Elsevier 2013

[6]Tarik S. Zakzouk Comparing text classifiers for sports news Elsevier 2012

[7]Ngoc Phuong Chau, Viet Anh Phan, Minh Le Nguyen Deep Learning and Sub-Tee Mining for Document Level Sentiment Classification KSE 2016

[8]Ivo Danihelka et. al. Associative Long Short-Term Memory arXiv 2016