Literature Review Of Different Methods For Sentimental Analysis

Kashish P. Kothari[1]

[1]Department of Computer Science – University of Illinois at Urbana-Champaign
Urbana, IL, USA

{kkotha4}@illinois.edu

1. Introduction

In today's world where the data collection and analysis is a major trend to understand the business. Data which tells the story helps to generate a lot of revenue if utilized properly. Social media platforms like twitter which contains opinions and reviews about businesses, organizations, and government, it has become important to generate sentiments about those opinions. It is not possible to manually check each sentiment therefore different sentiment analysis techniques are used. In recent times due to surplus data collected, advanced methodology such as deep learning is being used for sentiment analysis. Process of analyzing and summarizing the opinions expressed in the user-generated data is known as Sentiment Analysis or opinion Mining which has become a very interesting domain for researchers[4]. We are going to analyze and compare 5 different approaches in this paper. Lexicon based method, Hidden Markov model-based method, semantic orientation based method, Machine Learning based method, and deep learning based method.

Sentiment classification which classifies the user sentiments from microblog and tweets uses Lexicon based method, Hidden Markov model-based method, semantic orientation based method, Machine Learning based method, and deep learning based method. All these approaches have evolved in time with increase in computational capacity. The latest development combination of NLP and deep-learning technologies resulted in producing the most efficient model for sentiment classification. Deep learning based approach is the most recent which is able to analyze the sentiments in precise way compared to previous approaches.

The second section of the paper will describe about the scope of the literature and the overview of methods that are going to be analysed. The third section will describe each methodology in detail. The fourth section will describe data used in each methodology followed by the evaluation metric and detailed analysis of each approach.

2. Literature Scope

We are going to analyze the five main types of approaches used for sentiment classification. These approaches were developed as the advancement in technology and computation increased.

2.1. Lexicon based approach [6]

This is the traditional approach where lexicons are used for analyzing and classifying the sentiments. In this approach, common lexicons are created as a bag of words. Lexicon based approach is formed based on assumption that sentiments can be derived by the sum of the of all the lexicons in the sentence[6]. This type of approach weighs heavily on the semantic orientation of the sentence[2]. Lexicons are created manually and they are divided into two types common lexicons and content-based lexicons[6]. “Sentiment calculation is the aggregation of the sum of the sentiment-bearing entities of the tweet. Entities can be text, emoticons, and hashtags”[6].

2.2. Hidden Markov based approach [3]

Hidden Markov Model(HMM) based approach is an advancement of the lexicon-based method. This method considers syntactic as well as semantic information about the sentence while classifying the sentiments. General Features that are considered for classifying sentiments are part of speech tags, the polarity of lexicons and n-grams structure[3]. They have also included the modeling of word transition using syntactic and sentimental information instead of simple word-based transitions [3]. Kim et al. [3], have considered Hidden markov model for modeling the transition pattern from the training sentences and accurately estimating the probability of the sentence.

2.3. Semantic orientation based unsupervised approach [2]

Semantic orientation of the sentence plays an important role in sentiment classification. Turney [2] have predicted the sentiments based on the average semantic orientation of the phrases having adjectives and adverbs.Simple unsupervised classification depends on the average semantic orientation score.

2.4. Machine Learning based approach ([4],[8])

Machine learning techniques are of two types supervised learning method and unsupervised learning method. Sentimental classification falls under supervised learning method when data is trained on manually annotated sentiments. “The dataset might be boisterous and subsequently should be pre handled utilizing various Natural Language Processing (NLP) techniques”[4]. Generally, relevant features for sentiment analysis are extracted and finally the model is trained and tested on unseen data[4]. In this method, various machine learning algorithm such as Naive bayes, and SVM(support vector machine) are used.

2.5. Deep Learning based approach [5],[9]

Using deep learning it is possible to train the complex model on a huge dataset. Explaining the learned features is difficult in this type of approach. They have used Convolution neural network for feature extraction from text and Long short-term memory (RNN) for low dimension representation and sequentially processing the sentence. This approach is the most advanced method which learns the most complex features that other methods fail to learn.

3. Methodology

3.1. Simple Lexicon based approach:

Palanisamy et al. [6], describes lexicons into two main type, common lexicons, and content-based lexicons. Common lexicons include the sentimental words(with positive and negative sentiment score), negation words(reverses the polarity of sentimental words) and blind negation words such as “needs” in “this product needs to be better”. Content base lexicons are the words related to the content or product whose sentiment is described. Before applying sentiment classification algorithm Palanisamy et al. [6], have used few preprocessing techniques such as Part Of Speech(POS) tagging, stemming for getting stem words, emoticon detection and hashtag detection.

According to Palanisamy et al. [6], sentiments are classified based on the types of lexicons words that are present in the sentence. After preprocessing each word of the sentence is fed into the algorithm. Firstly, blind negation words are detected. if blind negations words are found than algorithm straight away gives negative sentiment as the classification result. If blind negation words are not found then algorithm looks for the sentimental lexicons with from the sentence along with sentimental score(+1 for the positive sentiment and -1 for the negative sentiment) . When a sentimental word is detected, it looks for the negation word in the range of 2 positions around that sentimental word. If negation word is found than the polarity value of a sentiment score is reversed. Sum of sentiment score in a sentence is mapped as the sentence score. Finally sentence is classified into positive, negative or neutral based on it’s score.

“A lexicon-based approach is very simple, viable and practical approach to Sentiment Analysis of Twitter data without a need for training”[6]. Results of Lexicon based method depends on the type of the lexicon it utilized.

3.2. HMM based approach:

The previous method just focused on the lexicons or semantic representation. This method takes into consideration syntactic representation along with semantic information by using hidden markov model(HMM).in [3], they have focused on groups of words that have similar roles from the syntactic and sentimental perspective. For example, “good” and ”nice” can be grouped with positive polarity as positive adjective into Sentimental information group(SIG)[3]. SIG is the group of word that consists of sentimental as well as syntactic representation.

For grouping unigrams into SIG, [3] have considered feature sets. These feature sets reflect the sentimental as well as syntactic information. Syntactic information includes part of speech tags and sentimental information consist of three types positive, negative and neutral. Overall Kim et al. [3] have created 76 SIG as some of the groups such as negation words don’t have sentimental information. SIG is divided into two main types, positive SIG and negative SIG based on it’s sentiments[3]. Some of the examples of SIG are a positive adjective, positive verb, and negative verb. The sentimental score of the word was identified using SentiWordnet if lexicons were not present than it was calculated by finding synsets of the word from Wordnet than looking at all those words into sentiWordnet for the score using the formulae:

$$sentiScore\left(w\right) = \frac{\sum_{v \in synset(w)} sentiScore_{SWN}(v)}{|synset(w)|}.$$

The words having positive score have positive sentiments and having negative score had negative sentiments.Final Model is divided into two types of classifier negative and positive classifier. Positive polarity SIG's were used for positive classifier and negative polarity SIGs for a negative classifier. Each word can belong to multiple SIG's for example "love" can be represented as either a positive noun or a positive verb[3]. Therefore each word is expressed as a feature vector representing different SIG's. Lastly, these feature vectors are then fed into HMM Classifier[3]. In HMM SIG are used as hidden state and initial emission probability is calculated using following formulae:

$$P\left(o_t = v_k \mid q_t = SIG_j\right) = P\left(v_k \mid SIG_j\right).$$

"In the equation, O_t and Q_t are the symbols and the state at t, respectively, V_k is the k-th unigram, and SIG_j is the j-th SIG"[3]. Transition probability is randomly initialized. Training is done using the Baum-Welch algorithm. After the HMM positive and negative classifiers are trained, the sentence is passed to both the classifier. Polarity with maximum likelihood is chosen as the classification for the input sentence[3].

3.3. Semantic Orientation based approach:

In this approach, the algorithm extracts the adjective and adverb from the sentences. Adjectives are described as a good indicator of subjectivity but they do not have enough context for semantic orientation[7]. In order to get the context, it extracts two separate words one for the context or subject and another word one can either be adjective or adverb. The first step consists of Part of speech tagging in order to extract adjective or adverb from the sentence. After POS step, Turney [2], look for the predefined pattern of the tags for the two consecutive words. Some of the patterns include JJ(adjective) followed by NN or RB(adverb) followed by an adjective and the following third consecutive word should not be NNS(noun). All identified patterns from the sentence are considered as identified phrase.

Semantic Orientation(SO) is calculated based on the PMI-IR calculation for the phrases identified in the above steps. PMI stands for pointwise mutual information and it is defined by the following formulae:

$$\mathrm{PMI}(word_1, word_2) = \log_2\left[\frac{\mathrm{p}(word_1\ \&\ word_2)}{\mathrm{p}(word_1)\ \mathrm{p}(word_2)}\right]$$

"p(word1 word2) is the probability that word1 and word2 co-occur. If the words are statistically independent, than the probability that they co-occur is given by the product of p(word1) and p(word2)"[2]. PMI is an indication of the statistical dependency of the words. IR stands for information retrieval for finding the words that are presented near the given sentiment expression. Turney [2],calculate semantic orientation by following formulae:

SO(phrase) = PMI(phrase, "excellent") - PMI(phrase, "poor")

Turney [2] have used excellent and poor as the reference words because in the standard rating system poor is referred to one star and excellent is refer to five stars. Higher sentimental orientation value means the positive association with excellent and lesser sentimental score indicates a negative association with excellent. In the final step, Turney

[2], calculate average SO value for all identified phrases from the review sentence. If SO is positive then a review is classified as recommended otherwise it is classified as not recommended[2].

3.4. Machine Learning based approach:

This Approach consists of various machine learning techniques that are effective for classifying the sentiments. predefined features are obtained from the document. Based on the predefined features, n_i(d) represents the number of times feature i appeared in document d [4]. Lastly, feature vectors are created from the above obtained values. These features are fed into the machine learning model and classification output is obtained. Features defined are based on the human introspection such as the lexicons humans observe while classifying the sentiments.Three standard machine learning algorithm are selected based on its previous performance on text categorization[4]. Naive Bayes Classification, Support Vector Machine and Maximum Entropy Classifier.

Naive Bayes classifier follows the bayes rule where class C = arg max_c P(c|d),

$$P_{\mathrm{NB}}(c \mid d) := \frac{P(c)\left(\prod_{i=1}^{m} P(f_i \mid c)^{n_i(d)}\right)}{P(d)}.$$

f_1,f_2,...,f_m are the m features used for the classification. P(c) and P(f|c) is trained using add one smoothing [4]. It is the most simple machine learning approach with the assumption that each feature are independent[8].

Maximum Entropy Classifier is another simple probabilistic approach which works better than naive bayes[4]. Maximum entropy tries to learn weight vectors for different feature distribution in order to classify the sentiment classes." iterative scaling algorithm was used for parameter training, together with a Gaussian prior to prevent overfitting"[4]

SVM are the most effective among the three algorithms in the text categorization[8]. SVM fits non linear data distribution very well because of its reliability on maximizing the margin rather than depending on probability[8]. SVM consist of support vectors which helps algorithm to learn separable boundaries thus maximising the distance from the support vectors of each class[4].

$$\vec{w} := \sum_j \alpha_j c_j \vec{d_j}, \quad \alpha_j \geq 0.$$

d_j is a document such that α_j is the support vectors helping to learn weights w for class C_j[4].

There are many advance algorithm such as decision trees which works on information gain helps to improve the result[8] . Machine learning results depend on the feature provided to the model for training. Feature engineering could lead to improvement in accuracy as well as the model capacity to understand complex sentences. Some other features may include POS+unigram, bigrams,unigrams + bigrams, unigrams + position and word embedding[4].

3.5. Deep Learning based approach:

Word embedding is an important step before applying Convolution Neural Network(CNN) in deep learning. Word embedding transfer the word into vector of d dimension[9]. Wang et al.[5], have used word2vec for converting words into the vector. Matrix is given by the dimension [d,V] where d is the dimension of vector and V is the vocabulary of words. The final sentence is represented by the S=[w1,w2,w3,w4....].

Convolution layer is applied after the word vectors is given as an input for specific window size in the sequence sentence[5]. Various filters are used for abstracting the different features from the possible windows of words. After the convolution filter, max pooling layer is applied to pool out most important features[9]. There is two type of pooling layer max pooling and average pooling in most of the cases max pooling is used. "The features generated from convolution and pooling operation can be viewed as advanced features like n-grams"[5]. For classification [5] have used RNN (LSTM) for processing sequential relations and learning long dependencies between the words in the sentence as shown in figure 1.

LSTM is long short term memory network which learns the relations between the context as well as words sequence[9]. At every time t, there is an input given to the network which outputs some probability. LSTM can be considered of multiple networks connected and each network represents a single word. LSTM have a hidden layer which gets passed across each time step and finally, output and loss can be calculated from the last time step[5]. "LSTM unit keeps the existing memory c_t at time t. The input at time t is x_t, h_t-1, c_t-1, the output is h_t, c_t"[5].

$$\begin{aligned}
i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\
f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\
o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\
g_t &= tanh(W_g x_t + U_g h_{t-1} + b_g) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot tanh(c_t)
\end{aligned}$$

In the above equations "sigmoid denotes the logistic sigmoid fuction. The operation $\odot$ denotes the element-wise vector product. At each time step t, there are an input gate i_t, a forget gate f_t, an output gate o_t, a memory cell ct and a hidden unit h_t. h_o and c_o can be initialized to 0 and the parameters of the LSTM are W,U, b"[5].

After RNN the last layer is the fully connected layer which outputs the probability of the sentimental classes and calculates the loss using softmax function. The formula for softmax loss is given by

$$\hat{P_i} = \frac{exp(o_i)}{\sum_{j=1}^{C} exp(o_j)}$$

P is represented as probability o_i represents the particular sentiment class. Summation of o_j represents all the sentimental classes value.

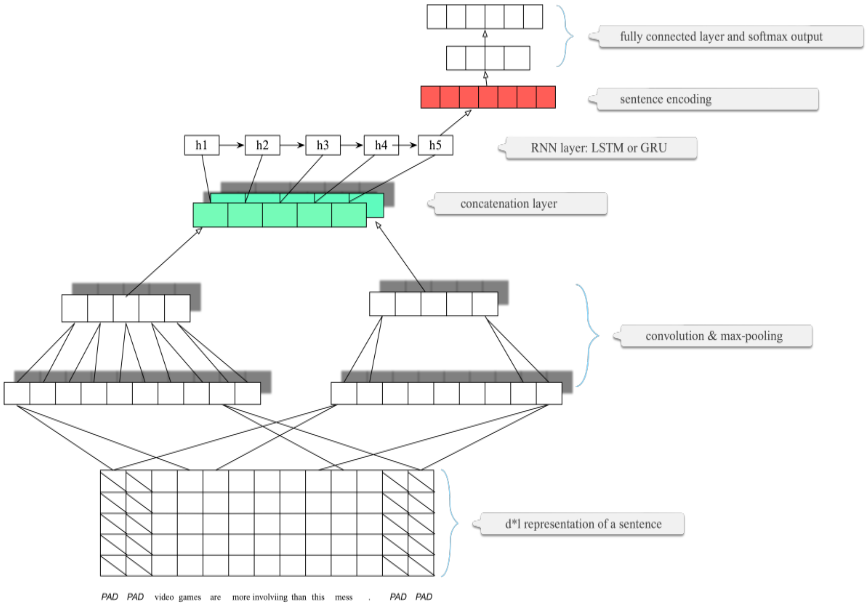
fully connected layer and softmax output
sentence encoding
h1
h2
h3
h4
h5
RNN layer: LSTM or GRU
concatenation layer
convolution & max-pooling
d*l representation of a sentence
PAD PAD video games are more involviing than this mess . PAD PAD

Figure 1. CNN-RNN

4. Data Collection and Pre-Processing

4.1. Lexicon based method [6]:

The real time tweets was used for training, 9411 subjective expressions were identified as sentimental lexicon from those tweets. preprocessing methods such as POS tagging,stemming,Emoticon detection, Exaggerated word shortening and Hashtag detection was used.

4.2. HMM based method [3]:

Health care reform dataset extracted from the twitter. 839 tweets was used for training set, 834 was used for validation set and 839 tweets was used for test set. Data preprocessing steps include separaitng username,url and emoticon from the tweets before training.

4.3. semantic orientation based method [2]:

Data was extracted from epinions a customer review website. 410 reviews was randomly sampled from 4 different domains. 170 are of negative sentiment and 240 were from positive sentiment. “The algorithm achieves an average accuracy of 74%, ranging from 84% for automobile reviews to 66% for movie reviews”[4].

4.4. Machine Learning based method [4]:

Data used was from “Internet Movie Database (IMDb) archive of the rec.arts.movies.reviews newsgroup.3” [4]. Selected reviews where the sentiments was expressed either in the form of score or stars. Limitation of 20 reviews per author per sentiment. Collected reviews from 144 authors. Prep-processing steps includes calculating unigrams, bigrams and part of speech tagging as a part of the feature engineering.

4.5. Deep Learning based method [5]:

Data involved was taken from three sources. Movie review with one sentence per review. Stanford Sentiment treebank which was an extension of movie review dataset but having five kinds of labels very negative, negative, neutral,positive and very positive.Third one was from sentiment treebank but with binary classes. Results were compared on three datasets with different models(LSTM,GRU) and hyper parameters. Word2vec is the only pre-processing step for deep learning approach

5. Evaluation Measures

As the sentimental analysis is a classification technique most commonly used measure for evaluating the model are:

$$Accuracy = \frac{number - of - correctly - classified - sentiments}{total - number - of - predicted - sentiments}$$

$$Precision = \frac{number - of - correctly - classified - class}{total - number - of - predicted - class}$$

$$Recall = \frac{number - of - correctly - classified - class}{total - number - of - actual - class}$$

$$F - score = \frac{2 * Precision * Recall}{Precision + Recall}$$

6. Analysis

In the lexicon-based method [6] we saw how the meaning of lexicon can be used in order to obtain the classification of the sentiments. This is a very simple approach, in a practical scenario just knowing the meaning of lexicon is insufficient for accurate classification. Another disadvantage of this method is that it did not consider the senses of the lexicon which plays an important role while defining the sentiments. The inclusion of relations between senses and word sense disambiguation techniques such as lesk algorithm in [10] as a feature would have increased the performance of the classifier. In HMM and SIG based approach, we saw that it considers syntactic relations as wells as sentimental features along with the word transition in HMM. The main advantage of this method is that it takes syntactic features such as part of speech tag while modeling the sentimental transition pattern. It starts with SIG which is fed to HMM. SIG features make robust HMM transition thus giving the better accuracy. According to [3] SIG with HMM outperforms than some machine learning model such as SVM and naive Bayes.

The third method [2], was based on the semantical orientation. It is similar to the lexicon-based method but it includes the adjective and adverbs to get the phrases. Thus finding the semantic orientation of phrases using PMI-IR and lastly averaging the phrase orientation for classifying the sentiments. Semantic orientation alone is not sufficient for obtaining good accuracy it requires additional features for classifying. The advantage of this method is that being an unsupervised approach saves a lot of time in an annotation of the training sentences. Being a simple method it is easy to understand compare to complex algorithm like deep learning. Sentimental orientation alone is not sufficient for producing an accurate result but when provided as a feature to the complex supervised algorithm than it makes the classification model more robust.

Machine Learning approach is based on a supervised classification where data is divided into train and test set. As we have seen Lexicon based methods require a lot of computation thus it is a time-consuming process. Machine learning is a scalable,fast and reliable technique[8]. Machine learning techniques learn the complex relation from the features and generate an output. Feature engineering is the most important step in the machine learning as the classification result depends on the features provided to the model to learn. SVM performs better compared to traditional naive Bayes and maximum entropy classifier in case of text categorization. Deep Learning approach is the latest and advanced technique used in multidisciplinary fields especially in NLP. CNN and RNN approach helps to learn a lot of complex features from raw input without explicitly providing features to the model. Pooling layer in CNN can learn the local features and relations while RNN model like LSTM and GRU can learn long-term dependencies and global features. Both combined together produce an excellent result in compare to all traditional methods. Deep learning is scalable and classifies huge data sets which traditional methods such as lexicon based are incapable.

7. Bibliography

[1] [Xue and Li 2018] Xue, W., and Li, T. 2018. Aspect based sentiment analysis with gated convolutional networks. In ACL, 2514–2523.

[2]Peter Turney.2002 Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th annual meeting on association for computational linguistics.

[3]Kim, Noo-Ri, Kyoungmin Kim, and Jee-Hyong Lee.2018 ”Sentiment Analysis in Microblogs Using HMMs with Syntactic and Sentimental Information.”International Journal of Fuzzy Logic and Intelligent Systems 17.4 (2017): 329-336 ScienceCentral. Web. 22 Oct.

[4] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?:sentiment classification using machine learning techniques. in Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2002). 2002.

[5] X. Wang, W. Jiang, and Z. Luo, ”Combination of convolutional and recurrent neural network for sentiment analysis of short texts,” in Proc. COLING, 2016, pp. 2428–2437

[6]Prabhu Palanisamy, Vineet Yadav, Harsha Elchuri,2013 “Serendio: Simple and Practical lexicon based approach to Sentiment Analysis”, Serendio Software Pvt Ltd.

[7]Hatzivassiloglou, V., & Wiebe, J.M.2000. Effects of adjective orientation and gradability on sentence subjectivity. Proceedings of 18th International Conference on Computational Linguistics. New Brunswick,NJ: AC.

[8] Anuja Jain, Padma Dandannavar,2016. ”Application of machine learning techniques to sentiment analysis”, 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), pp. 628-632.

[9]A. Hassan and A. Mahmood,2018. ”Convolutional recurrent deep learning model for sentence classification,” IEEE Access, vol. 6, pp. 13949–13957.

[10]Hockenmaier,Julia.CS447: Natural Language Processing https://courses.engr.illinois.edu/cs447/fa2018/Slides/Lecture18.pdf