ITC 4/49

Information Technology and Control
Vol. 49 / No. 4 / 2020
pp. 564-582
DOI 10.5755/j01.itc.49.4.25350

Movie Aspects Identification Model (MAIM) for Aspect Based Sentiment Analysis

Received 2020/02/22 | Accepted after revision 2020/10/13

http://dx.doi.org/10.5755/j01.itc.49.4.25350

**HOW TO CITE:** Mir, J., Mahmod, A. (2020). Movie Aspects Identification Model (MAIM) for Aspect Based Sentiment Analysis. *Information Technology and Control*, 49(4), 564-582. https://doi.org/10.5755/j01.itc.49.4.25350

# Movie Aspects Identification Model (MAIM) for Aspect Based Sentiment Analysis

**Jibran Mir, Azhar Mahmood**

Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST), Faculty of Computing and Engineering Sciences, Computer Science Department, Islamabad, Pakistan

**Corresponding author:** dr.azhar@szabist-isb.edu.pk

Aspect Based Sentiment Analysis techniques have been widely applied in several application domains. During the last two decades, these techniques have been mostly developed for the domain of product and service reviews. However, very few Aspect Based Sentiment Techniques have been proposed for the domain of movie reviews. In contrast to most studies that focus on movie specific aspects such as Script, Director, and Actor, this work focus on NER (Named Entity Recognition) in order to find out entity-specific aspects. Consequently, MAIM (Movie Aspects Identification Model) is proposed that can extract not only movie-specific aspects but can also identify Named Entities (NEs) such as Person Name and Movie Title. The three main contributions in this paper are (i) identification of infrequent aspects, (ii) identification of NEs, and (iii) identification of N-gram opinion words as an entity. MAIM is implemented using BiLSTM-CRF (Bidirectional Long Short-Term Memory – Conditional Random Field) hybrid technique and tested on movie reviews dataset. The results showed a precision score of 89.9%, recall of 88.9%, and f1-score of 89.4%. The results of the hybrid model are compared with the baseline models i.e., CRF (Conditional Random Field) and LSTM-CRF (Long Short-Term Memory – Conditional Random Field) and shown hybrid model outperforms both models in term of precision, recall and f1-score.

KEYWORDS: Movie A pplication domain, Explicit Aspects, Named Entity Recognition, Opinion words, Aspect Pruning and Aspects identifications.

## 1. Introduction

The most intelligent beings in this planet earth are humans since, naturally, they possess cognitive or decision making ability [22]. In addition, psychologists have categorized the decision making ability into further two categories, namely, rational decision making and irrational (emotional) decision making [33]. However, few studies [8, 33] emphasize that the irrational or emotional entities have more influence than rational or reasoning in making decisions. Therefore, the most striking factor in decision making is emotions or sentiments. One crucial question arises here which is: where to find emotions or sentiments in order to make efficient decisions? The obvious answer is the social circle in any country or city. However, in the twenty first century this social circle has been turned into digital social circle. In other words, this is the age of digital or the internet world.

The participation of a huge number of people in social media has been made possible by participatory web or web 2.0 [10] which motivates common users to produce their own knowledge. In addition, the user generated knowledge is unstructured dataset and according to data scientists [29] this is the most challenging dataset, in order to extract relevant knowledge out of it. Unstructured content is the most favorable type of content produced by the common user since it allows the user to express his/her feeling/sentiment regarding an entity without any structural restriction. For example, Ecommerce and social networking websites provide a platform for common users [40] to write reviews (unstructured content) about a product or social event, thus allowing them to express their feeling/sentiment regarding the product being reviewed. Business owners then evaluate their product performance by analyzing these reviews. Moreover, customers too use these reviews for a perfect purchase decision. This thus also becomes a source of knowledge.

Since data volume is increasing day by day, it is humanly impossible to read and analyze every customers' or business owners' review for sentiment exploration. However, information scientists have proposed sentiment analysis techniques that can extract sentiments from a large number of reviews. These sentiment analysis techniques can be categorized into three levels, these levels being Document-level, sentence-level and aspect level.

The document-level sentiment analysis simply focuses on whether the overall review is negative or positive. It works on the assumption that any review that is being analyzed, talks about only one single entity or dimension [29, 40]. Sentence-level sentiment analysis, on the other hand, explores more intense sentiments than the document-level analysis. The sentence-level sentiment analysis focuses on subjectivity analysis. It distinguishes the sentences, within a review, into subjective (negative and positive) and objective (factual) sentences [55]. In contrast to both of these, the most fine-grain method of sentiment analysis is "aspect-based sentiment analysis". Aspect-based sentiment analysis is important since neither the sentence level and nor document level sentiment analysis performs an aspect level sentiment analysis, [40]. For example, if a review says, "The battery life of iPhone is very nice", the sentence level sentiment analysis would suggest that the whole sentence is subjective, since it contains opinion. It would not provide any detail of the iPhones aspects. On the other hand, aspect level sentiment analysis would explore "battery life" as aspect and "very nice" as a positive sentiment word. Hence, the aspect level sentiment analysis explores an entity's various aspects with related sentiment words.

The application area of aspect level sentiment analysis shapes the working body for aspect based sentiment analysis models; for instance, preprocessing, dataset annotation, selection of machine learning techniques and then the performance evaluation criteria. While proposing a model for aspect level sentiment classification, it is crucial to probe the application domain carefully. Data scientists attempt to extract relevant information from a large available noisy dataset. However, this task becomes more difficult when an application domain changes. In addition, the difficulty level of data extraction varies from application to application. For example, a cellphone application domain contains very simple aspects such as "price", "weight", "performance", "sound quality" [14]. The reviews related to this application domain are thus less noisy and very simple, since the commentators themselves are very straightforward and comment only on the aspects they like or dislike. As a result, the current ABSTs (Aspect Based Sentiment

Analysis Techniques), which have been designed for this product application domain, are not very complex and have successful performance accuracy.

In contrast, a service application domain, such as the tourism domain, is more complex and as such, the reviews related to this application are noisy. The reason is that the commentators, for this domain, are often keen to define a proper story of his/her trip. The aspects related to this domain are “food”, “atmosphere”, “staff”, “room” etc. and [39] the commentators usually write in indirect speech rather than direct speech. Naturally then, the models that have been designed for this particular application domain are more complex. Similarly, the movie application domain is more complex than both the product and the tourism application domains. The reason is that this application domain contains co-resolution reference and name entity recognition (NER) problem along with too much noise and indirect speech. The models that would be designed for this domain would then be even more complex than the models designed for either the product or the tourism application domain. This shows that the complexity of an ABST (Aspect Based Sentiment Analysis Technique) depends on application domain for which it has designed.

Several ABSTs (Aspect Based Sentiment Analysis Techniques) have been proposed for different application domains. However, most of the application domains are product [3, 21, 60] and service [15, 39]. Moreover, very few ABSTs are on movie application domain [7, 42].The ABSTs which have been proposed for movie application domain are unable to satisfy the requirement of intense or aspect level sentiment analysis. Subsequently, the movie application domain is more complex than product and service application domain and it contains NER (Named Entity Recognition) problem, indirect speech, and noisy content.

In this article, ASBT has been proposed and implemented for movie application domain. The following are the main contributions.

1. MAIM (Movie Aspects Identification Model) extracts movie aspects without any discrimination of frequent and infrequent aspects by using local context. In addition, these infrequent aspects cannot be ignored, though, they are associated with opinion.
2. The movie application domain contains NER problem and previous Movie ABSTs [7, 9, 30, 42] ignored the identification of NER problem. The MAIM is using BiLSTM-CRF (Bidirectional Long Short Term Memory – Conditional Random Field) hybrid NERC (Named Entity Recognition and Classification) technique, for the identification of Person Name and Movie Titles in movie application domain.
3. MAIM not only identifies NOUN based entities such as Person Names and Movie Titles, but, also identifies ADJECTIVE or ADVERB based opinion entity.

## 2. Literature Review

### 2.1. Aspect Identification Techniques for Movie Application Domain

Anand et al. [7] has chosen the movie reviews as social application domain. This study has two main contributions filtering out irrelevant plot and identification of aspects and opinion words. However, the way they have filtered the review text, results in having missing information. For aspect identification they are using dependency parser and to determine the polarity of the associated opinion word they are using WordNet. A list of movie aspects i.e. (acting, directing, screenplay, sound effect and music, story, visual effects and movie on the whole) have already defined and the identified aspects clues will be clustered to one of the aspects in the list. Consequently, the proposed model completely ignored the titles of the movies and persons that are also associated with opinion.

Similarly, Parkhe et al. [42] has proposed an aspect based sentiment analysis model for movie reviews. The main contribution of this study is the clustering of review sentences to the already identified aspects, for instance, screenplay, music, acting, plot, movie and direction. After that they determine the polarity of each sentence. Nevertheless, the proposed model is restricted to six basic aspects only and ignores Person Names and Movie Titles which are also associated with opinion. In addition, the proposed model is unable to identify other aspects which are not in the seeds set. Apart from having aspect level sentiment classification, majority of the opinion mining techniques have been proposed for document level. Therefore, Maas et al. [37] has proposed a document level sentiment analysis technique. The main contribution of this research is the classification of movie reviews by using Unigram, Bigram, Trigram or amalgamation all of them.

Similarly, Tripathy et al. [54] has applied a hybrid technique on dual IMDb datasets (Association for Computational Linguistics Internet Movie Database, Cornell Movie-review). The contributions of this research are twofold. First, they have selected sentiment features by using SVM (support vector machine) classifier and then incorporated these features into ANN (Artificial Neural Networks). The proposed model performs document level sentiment analysis. In contrast to document level sentiment analysis, Ahmed et al. [4] has proposed a technique for sentence level sentiment analysis and this technique has validated its performance by using aclIMDb (Association for Computational Linguistics Internet Movie Database) dataset. Another document level sentiment analysis has been proposed by Panda [41]. Moreover, its main contributions are pre-processing to refine the dataset for feature selection, prune the feature sets and then apply algorithm-SGNB (Sparse Generative Naive Bayes) for sentiment classification. For the validation of the proposed model they have used aclIMDb (Association for Computational Linguistics Internet Movie Database) dataset. However, these sentiment analysis techniques are incapable of identifying NER problem and the opinions that have been expressed for them. Therefore, there is a need to incorporate NER (Named Entity Recognition) technique in aspect based sentiment analysis for movie application domain.

#### 2.1.1. Named Entity Recognition Techniques

Named-Entities are also treated as aspects in some domains, for instance, in the movie application domain. The Person Names and Movie Titles are widely used in movie application domain. Moreover, these entities are associated with opinions. Therefore, it is important to identify opinion rich named entities. According to our knowledge, few named entity techniques offered opinion identification along with name-entities. Regarding this, Şeker et al. [51] has proposed a NER (named entity recognition) technique for Turkish new articles. The entities this research has identified are person names, time and number. The contribution of this research is to identify entities from morphological rich language for instance Turkish language. However, no opinion identification method has been defined in the proposed research. In contrast, Mir et al. [27] has proposed movie NER technique and it also provides opinion identification. The main limitation of the proposed research is that its annotation phase generates noisy labelled dataset. In addition, it did not provide any approach to remove noisy aspects. Finally, the proposed technique solely depends on CRF technique; however, latest NER techniques proved that the CRF alone will not perform efficiently.

In the same way, Lample et al. [32] has proposed a hybrid NER technique for the identification of person, location, organization and other entities. This study states that the previous models heavily dependent on handcraft linguistic features and domain specific knowledge. The proposed model uses a hybrid neural network technique BiLSTM-CRF (bidirectional long short term memory – Conditional Random Field) and provides better performance than baseline models. Similarly, Do et al. [18] has also proposed a hybrid NER technique for the identification of named entities in plant molecular biology dataset. The proposed technique identifies genes, proteins and phenotypic Traits as named-entities. However, it is incapable of identifying opinion words.

### 2.2. Aspect Identification Techniques for Different Application Domains

Rana et al. [46] has asserted that the NLP (Natural Language Processing) tool such as POS (Part of Speech) tagger is incapable of correctly tagging the part of speech for customer reviews because the review writer of a product does not follow any grammatical rules. Therefore, to overcome this inefficiency of POS tagger the author has proposed its own linguistic rules to extract the potential aspects. However, on one side the author highlighted POS tagger's incapability of correctly tagging the customer reviews, on the other side, he used opinion lexicon (which is also a linguistic tool) to verify whether the wrongly tagged adjective is a NOUN/NOUN PHRASE or not. However, opinion lexicon also suffers the same inefficiency which has been identified by the research [46]. In addition, to filter out irrelevant aspects from the customer reviews the author has proposed two pruning techniques frequent filtering and semantic filtering. However, the frequent filtering also filter outs infrequent but relevant aspects. Apart from aspect identification, the proposed model is domain specific.

For instance, if the proposed model would have been applied to other domains such as movie or tourism domain, the performance could go down fairly, since the proposed model only identifies frequent and product application domain aspects. However, the movie application domain contains name entities such as Person Name and Movie Title, as an aspects of movie application domain. Therefore, the proposed model is unable to identify NE (named entity).

Ikram et al. [26] has proposed an aspect level sentiment analysis technique for selecting a research paper which has more positive aspect level sentiments. This research has followed the usual phases of aspect based sentiment analysis that are (1) aspect identification and (2) opinion determination. In order to identify aspects from citation corpora, the author has used linguistic patterns, word frequency, synonyms and linguistic rules. The research considers a NOUN or NOUN PHRASE as aspect if it is most occurring aspect in the corpus. Therefore, if an aspect is infrequent and holds an opinion, it will be ignored. The proposed research is domain specific and incapable of extracting aspects from other application domains. For instance, the movie and tourism application domains contain named-entities as aspects; however, the proposed research is unable to identify these entities, since it is an unsupervised technique.

Al-Smadi et al. [5] has presented an aspect based sentiment analysis model by using long short-term memory deep neural networks for tourism Arabic reviews. The major contribution of the proposed research is to use deep learning and CRF approaches to identify aspects of tourism and their sentiment determination. The proposed research is supervised and it requires plenty of labelled data. The author has described very few annotation information that are essential for supervised techniques. It is a domain specific model since the model has been trained on Arabic tourism domain. The annotated dataset has been labeled according to tourism application domain feature set. Hence, the proposed research is incapable of identifying NE (named-entity) in movie application domain.

Salas-Zárate et al. [47] has proposed an aspect based sentiment analysis technique for tweets on Diabetes. The core of this research is to use Diabetes ontology for aspect identification. A semantic annotation module has been proposed which comprises of Stanford NLP and Diabetes ontology. It defines the relationship between different concepts in Diabetes domain. The proposed research is very specific to Diabetes domain, since it depends solely on ontology. The proposed model, when applied to product, tourism or any other application domain is incapable of identifying accurate aspects, since Diabetes ontology is unable to define relation of concepts in other domain. In contrast, the Al-Smadi et al. [6] has proposed an aspect level sentiment technique for hotel reviews in Arabic language. For features extraction, the author used Arabic NLP tool MADAMIRA (Morphological Analysis and Disambiguation of Arabic) and using this tool morphological, syntactic, semantic features has been extracted. The proposed model asserted that it has proposed three types of identifications which are Aspect Category Identification, Opinion Target Extraction and Polarity Identification.

However, the proposed research is following a supervised technique which requires efficiently annotated dataset and the research lacks in defining any type of annotation process. This research also identifies named-entities by using Arabic NLP tool. However, this tool will be incapable of identifying named-entities in IMDb reviews. Similarly, another aspect category classification technique has been proposed by Schouten et al. [50]. This research has exploited association rule mining and frequency of co-occurrence data. The main contribution of this method is avoiding the supervised labeled training dataset and mining aspect level sentiment using unsupervised technique. However, the proposed research is incapable of identifying infrequent aspects and NEs, since it is an unsupervised technique and heavily depends on frequency of words.

In contrast, [58] the study has conducted research on Hate Crime Twitter Sentiment dataset in order to develop aspect based sentiment analysis technique. The major contribution regarding aspect classification is to use association rule mining with POS tagging patterns to identify single or multi-word aspect identification for Twitter reviews. However, the proposed research has been designed for short reviews, therefore, it will not perform efficiently on longer reviews. The reason is that the longer reviews contain more irrelevant data. In addition, the proposed research used an unsupervised method that works frequency of data, however, this will result in missing of infrequent but important aspects. Since it is an unsupervised model, it will also incapable of identifying named entities from reviews efficiently.

## 3. Critical Analysis and Limitations of ABSTs

The ABSTs have been implemented by using supervised learning, unsupervised learning, syntactic techniques, dictionary based or an amalgamation of these techniques [49] and the recent trend is, deep learning techniques. Scientists of sentiment analysis are still using each of these techniques, while keeping in mind their merits, demerits and application domain specific requirements. Unsupervised state-of-the-art ABSTs depend on frequency term, therefore, they only extract frequent aspects and ignores less but important aspects [12]. However, these techniques do not require heavy labeled dataset and require less computing time. In contrast [30] has used Deep Learning Convolutional Neural Network for aspect based sentiment analysis, but the annotation process is dependent on frequency based technique. Hence, this technique also identifies frequent aspects and CNN (Convolutional Neural Network) has been trained on this annotated dataset, it will also suffer from this problem. In addition, [38] is a supervised aspect based sentiment analysis technique, nonetheless, aspect identification process is based on TF-IDF (term frequency–inverse document frequency) technique, conversely, it is also dependent on occurrence of a term in a corpus.

Another limitation that exists, in several studies such as [43], [24], [44], [45], [27], is the consideration of irrelevant aspects as potential aspects. For instance, the linguistic tool such as POS tagger labels a word "lion" as NOUN and every NOUN is considered to be an aspect, therefore, "lion" is an aspect. However, in product application domain this is not a relevant aspect. Apart from these limitations, NER problem, which exists in service and movie application domain, has not been addressed in aspect based sentiment analysis techniques [13], [61]. One reason for this limitation is that the majority of the aspect based sentiment analysis techniques have been designed for product application domain, therefore, this limitation has been ignored. However, in service application domain location name and hotel names and in movie application domain Person Name and Movie Titles are considered as aspects of respective domains. Table1 shows some of the limitations of aspect based sentiment analysis.

**Table 1**

Limitations of Aspect Based Sentiment Analysis Techniques

| | |
|---|---|
| $L_1$ | Classifies only frequent occurring aspects and ignores the less occurring but important aspects [12, 30, 38]. |
| $L_2$ | Unable to provide any aspect pruning technique to filter irrelevant aspects [24, 43]. |
| $L_3$ | Unable to address NER (name entity recognition) for movie application domain [7, 9, 13, 30, 42, 61]. |

Recently, deep learning techniques have been employed in aspect level sentiment classification. Not surprisingly, their results are better than NLP and machine learning techniques [11]. Nonetheless, they require huge amount of training dataset and computational power. Apart from this, if an Aspect Based Sentiment Analysis technique is incorporating deep learning approach then it does not mean it will also identify infrequent aspect along with frequent ones. Unless or until dataset annotation is free from frequency based techniques [30]. $L_1$ points to the limitation that exists in unsupervised [12] or supervised [38] or even in deep learning if the annotation is based on frequency based techniques [30]. An example of $L_1$ is in product application domain is that the commentator often comments on the aspects such as price, battery life, design and brand or company. Therefore, the existing research targets the most occurring NOUN in the product application domain and ignore the aspects that are less occurring even they have strong association with sentiment. For instance, "Hawaii is the most favorable mobile in South Asia. Since, it is less in price, beautiful in design, fast in speed. But after one year of usage its speakers goes down." Now, if the most frequently occurring aspects in this comment are ***price, design and performance*** then, according to the current aspect based sentiment analysis techniques [59], [35], [52] these are aspects whereas the "speakers" as an aspect will be ignored regardless of a strong negative sentiment. In the explicit aspect identification the most stirring Part of Speech tag NOUN is considered to be as aspect. Conversely, not every NOUN/NOUN PHRASE is aspect, hence, the aspect level identification techniques only take frequently appearing noun. The driving factor behind this is the assumption that if 100 commentators comment on a particular mobile then probably they are talking about *price, design or performance*. Nonetheless, in reality they are skipping useful information which is also important for aspect based sentiment analysis.

To overcome this $L_1$ the investigation of reviews is important. Therefore, in this research we are using a hybrid method to extract explicit aspects that not only targets frequent NOUN/NOUN PHRASE but also infrequent ones as well. Since, the annotation process of movie reviews is not using any frequency based technique or any other assumption.

$L_2$ refers to the limitation that is an essential part of the aspect based identification process. Elimination of irrelevant aspects that are introduced by NLP tool, *POS tagger*, are important to tackle in aspect identification process for accurate results. The NOUN/NOUN PHRASE is considered to be as aspect, conversely, not every NOUN/NOUN PHRASE is a relevant aspect for a particular domain. For instance, in this sentence "Story of a man who has unnatural feelings for a pig" feelings and pig are NOUN but these are irrelevant words to the movie application domain. The state-of-the-art [43], [24], [44], [45], [27] studies have either ignored this matter or their aspect pruning techniques are not efficient. In order to overcome $L_2$ the proposed technique extracts NOUN/NOUN PHRASE, from the movie reviews, without consideration of its frequency the in corpus. Conversely, the irrelevant aspects will be eliminated by using semantic similarity matrix and SpaCy[1] is an NLP tool that provides modern technique for term similarity matrix.

$L_3$ highlights the existence of NER problem in movie application domain. The state-of-the-art aspect based sentiment analysis techniques have been mostly designed for product and service application domains. In product application domain NER problem does not exist, hence, the proposed solutions for product application domain do not facilitate the identification of NER problem. Contrastingly, the service and movie application domains both contains NER problem. In addition, none of the state-of-art ABSTs, which have been proposed for service and movie application domains, provide solution for NER problems [7, 9, 13, 30, 42, 61]. Since, this research is undertaking the investigation and development of ABST for movie application domain, therefore, it is important to handle NER problem (Person Names and Movie Titles) in aspect based sentiment analysis. The NER problem cannot be neglected because there is an opinion often associated with these name entities.

Previously, for the identification of Named Entities CRF (Conditional Random Field) was considered to be very efficient [31]. However, the successful use of deep learning techniques in NLP (Natural Language Processing) domain encouraged the information retrieval experts to use in identification of Named Entities from a corpus. Several studies [20, 28] employed CNN (Convolutional Neural Network) in document level sentiment analysis. However, according to a comparative study [57], CNNs is not suitable for NER problem, whereas, it reassures that the invariants of RNNs (Recurrent Neural Networks) are suitable for NER problem. Though NER problem has been extensively studied in information retrieval field [16, 18, 32] and it is in progress as the topic of research. Yet, there is a need to include these techniques into sentiment analysis domain.

The first challenging task in integrating NER technique is for the proposed model to have an annotated dataset for movie application domain. The most popular annotated dataset available is CoNLL (Conference on Natural Language Learning)2003 [1, 56]. This dataset was built from newswire articles and provided annotation for four entities (PERSON, LOCATION, ORGANIZATION and MISCELLANEOUS). However, the CoNLL (Conference on Natural Language Learning) 2003 annotated dataset's application domain is different, therefore, in this research an automated annotation technique has been provided for the annotation of movie application domain.

## 4. Proposed Movie Aspects Identification Model (MAIM)

MAIM consists of three main phases, however, the first two are more important phases such as annotation phase and training phase. Figure 1 depicts the three phases of movie application domain, for in-

1 https://spacy.io/usage/vectors-similarity

stance, annotation phase, training phase and identification phase, whereas, training and identification phases have been merged into one phase in Figure 1. The annotation phase includes linguistic features, orthographic features and movie gazetteers. In addition, this research has also exploited the capabilities of CRF and variant of Recurrent Neural Networks (BiLSTM-CRF). However, before applying these machine learning techniques on any dataset, the perquisite is the training or annotation of dataset. Therefore, the movie application domain needs to be annotated for the identification of movie titles, person names, movie specific aspects and opinion chunks.

## 4.1. Feature Set for Annotation of NER Dataset

### 4.1.1. Linguistic Features

The OpenNLP[2] tool for natural language processing has been used, this NLP software's tokenizer, POS tagger and the chunker has been used to annotate the unstructured movie reviews dataset. The input is an unstructured file but the output will be a structured five column file. Table 2 provides Algorithm I for converting unstructured text into a structured text which has been labeled with POS, Chunk, Title Case and IOB (Inside outside beginning) tags. The abbreviations for variables in Algorithm I are TK for Token, CK for chunk, CL for CaseLabel and iobT for IOBtag.

**Table 2**

Algorithm I Converts Reviews into sequence of tokens and labels POS, chunking and orthographic features to these tokens

**Input**: A set of unstructured text ***n*** movie reviews $R = \{r_1, r_2, \ldots, r_n\}$

**Output**: A set of five columns file set $Ff = \{Ff_1, Ff_2, ..., Ff_j\}$

```
for i ← 0 to n total number of reviev
while r_i read each line L to the end of file.
  TK_j ← MakeToken(L)
  POS_j ← POStagger(TK_j)
  CK_j ← Chunker(TK_j, POS_j)
  for j ← 0 to n number of Tokens
    caseLabel ← TagCase(TK_i)
    iobTag ← "O"
    Ff_j ← writeF(TK_j, POS_j, CK_j, CL, iobTag
  endfor
endwhile
endfor
```

The set of reviews $R = \{r_1, r_2, \ldots, r_n\}$ will be treated as input. The output will be a set of five columns files denoted by $Ff = \{Ff_1, Ff_2, ..., Ff_j\}$ where $Ff_1$ is a single file that contains five columns such as Token Word, POS, Chunk, Case and IOB. The outer-loop controls number of text files while inner loop controls the number of lines in each text file. Then each line is tokenized into sequence of token words denoted by $Token_j$ where *j* denotes the current token. Similarly, $j+1$ denotes the next token $Token_{j+1}$ and $j-1$ denotes the previous token $Token_{j-1}$. Therefore, $Token_j = \{token_1, token_2, token_3, ....., token_j\}$ represents the first column of file $Ff_i$. In the same way, $POS_j = \{pos_1, pos_2, pos_3, ...., pos_j\}$ represents the second column of file $Ff_i$. Conversely, $Chunk_j = \{chunk_1, chunk_2, chunk_3, ....., chunk_j\}$ represents the third column of file $Ff_i$. The method $MakeToken(\mathrm{L})$ takes line L from $r_i$ and outputs an array of $Token_j$. This $Token_j$ will be an input for the method $POStagger(Token_j)$ and the output will be an array of parts of speech $POS_j$ for each token word. Moreover, the method $Chunker(Token_j, POS_j)$ receives array of tokens and array of parts of speech as an input. In this case, the output will be array of chunk information $Chunk_j$. Moreover, there is another inner loop which iterates $j = 0$ to $N$ number of tokens. The method $TagCase(Token_j)$ takes current token as input and calculates its case information and outputs $caseLabel$. Meanwhile, IOBtag will be assigned letter "O". Finally, the method $Ff_j \leftarrow writeF(TK_j, POS_j, CK_j, CL, iobTag)$ takes five parameters as an input and produces a five column five $Ff_j$ file as an output. Figure3, demonstrates an excerpt of annotated sentence from movie corpus. The sentence has been tokenized and each token has its corresponding Part of speech, chunking, case and IOB information.

### 4.1.2. Orthographic Features

In writing text, orthographic features are very important, the writer uses these features to distinguish one word from the other words in a sentence. CoNLL-2003 Shared Task [48] uses orthographic features to generate annotated dataset for NERC (Named Entity Recognition and Classification). Similarly [2], uses orthographic features to identify NER (named entity recognition) in social media. In contrast [36] asserted that in biomedical application domain, orthographic features are very important in order to get the essential

2 https://opennlp.apache.org/

results. Orthographic features are usually prefix, suffix, word case and punctuation. To annotate the aclIMDb dataset the orthographic features have been used.

#### 4.1.3. IMDb Database

In order to remove irrelevant Person Names and Movie Titles, IMDb database will be used, because the Person Names are NOUN/NOUN PHRASE and not every NOUN/NOUN PHRASE is a relevant Person Names or Movie Titles. While extracting and applying linguistic patterns, irrelevant information is also extracted. To refrain irrelevant information, IMDb python application has been used to eliminate irrelevant NOUNS. For instance, to tag the Person Names all the token words that are labeled with NN or NNS POS information, Word Case is TC and chunking information is B-NP or I-NP will be extracted, without considering the frequency of each token word. However, this set of linguistic rules also mines irrelevant Person Names as well.

**Table 3**

The output of Algorithm II, which depicts that the linguistic rule labelled the correct bigram words as Person Name

| Word | POS | Chunk | Case | IOB |
|---|---|---|---|---|
| Vilmos | NNP | I-NP | TC | B-Person |
| Zsigmond | NNP | I-NP | TC | I-Person |

From Table 3, if the current word's POS label is NNP and its chunking label is B-NP or I-NP and its word case Label is TC then this token word would be tagged with Person Entity. Conversely, this rule also extracts irrelevant named entities. For example, Table 4 shows the irrelevant extraction of named entities. Orchestra Audience is not the name of the person but it has been extracted as named entity.

**Table 4**

The output of Algorithm II, which depicts the linguistic rule labelled the incorrect bi-gram words as person name

| Word | POS | Chunk | Case | IOB |
|---|---|---|---|---|
| **Orchestra** | NNP | I-NP | TC | B-Person |
| **Audience** | NNP | I-NP | TC | I-Person |

The imdbpy[3] (internet movie database python) software has been used to filter out irrelevant names from the annotated dataset. In the same way, Movie Titles have been tagged. However, this time bi-gram tokens have not been used, although, N-gram tokens have been extracted. Since movie titles are often more than two words, a similar rule to Algorithm II has been used to extract Movie Titles. The irrelevant N-gram words have been removed by using IMDb (internet movie database) database python application program. If an N-gram word exists in the database, then it is a Movie Title, otherwise, it will be discarded.

**Figure 1**

The Proposed Movie Aspect Identification Model (MAIM)

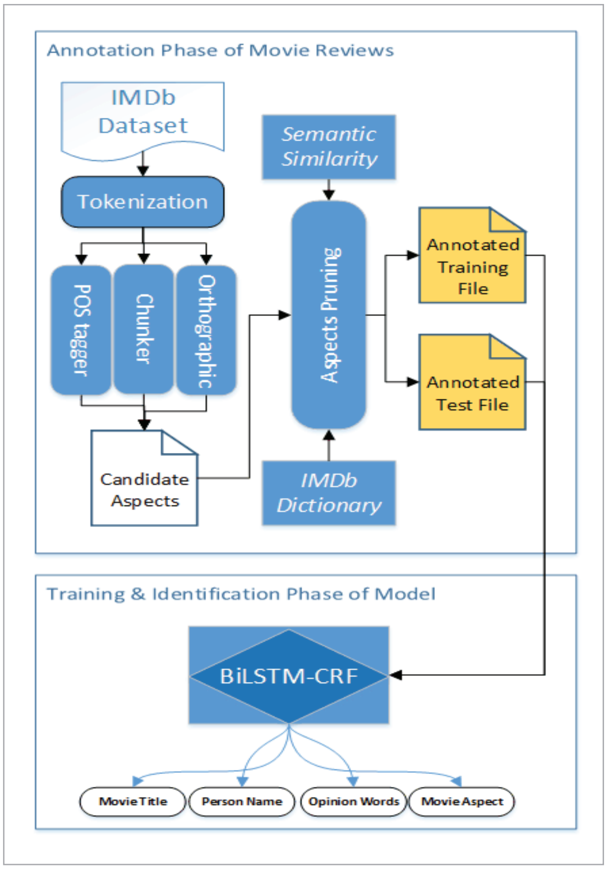



Table 5 shows Algorithm II, which takes a set of five column file $Ff = \{Ff_1, Ff_2, ..., Ff_j\}$ and provides a set of person and movie title annotated files $T = \{t_1, t_2, t_3, ..., t_n\}$. The line number 4's iterates $j \leftarrow 0$ to n total number of Tokens. In every iteration, current token word's *POS, CaseLabel and Chunk* will be conditionaly checked. If the condition satisfies the next token word's $POS_{j+1}$, $CL_{j+1}$ and $CK_{j+1}$ will be

3 https://imdbpy.github.io/

**Table 5**

Algorithm II Labels IOB scheme tagging for Named Entities in five column file Set $Ff = \{Ff_1, Ff_2, ..., Ff_j\}$

**Input:** A set of five column files $Ff = \{Ff_1, Ff_2, ..., Ff_j\}$
**Output:** A set of training files $T = \{t_1, t_2, t_3, ..., t_n\}$

*for* $i \leftarrow 0$ to n total number of four column files
  *while* $Ff_i$ read each line L to the end of file
   $TK_j, POS_j, CK_j, CL_j, IOBtag_j \leftarrow$ L
   *for* $j \leftarrow 0$ to n total number of Tokens
    *if* $(POS_j$ is NOUN$)^{\wedge}(CL_j$ is TC$)^{\wedge}(CK_j$ is B-NP$)$
     *if* $(POS_{j+1}$ is NOUN$)^{\wedge}(CL_{j+1}$ is TC$)^{\wedge}(CK_{j+1}$ is I-NP$)$
      $PersonName = Cancatinate(TK_j, TK$
      *if* $PersonName$ is in $imdbPython$
       $IOBtag_j \leftarrow$ B-PERSO
       $IOBtag_{j+1} \leftarrow$ I-PERSO
      *endif*
     *endif*
    *endif*
   *endfor*
  *endwhile*
*endfor*

**Table 6**

Algorithm III Labelling IOB scheme for general aspects in five column annotation files set $Ff = \{Ff_1, Ff_2, Ff_3, ....Ff_n\}$

**Input**:A set of five column files $Ff = \{Ff_1, Ff_2, Ff_3, ....Ff_n\}$ and twelve seed words $S = \{s_1, s_2, s_3, ...., s_n\}$
**Output**: A set of training files $T = \{t_1, t_2, t_3, ..., t_n\}$

*for* $i \leftarrow 0$ n total number of four column files
  *while* $Ff_i$ read each line L to the end of the file
   $TK_j, POS_j, CK_j, CL_j, IOBtag_j \leftarrow$ L
   *for* $j \leftarrow 0$ to n total number of Tokens
    *if* $(POS_j$ is NOUN$)^{\wedge}(CL_j$ is LC$)^{\wedge}(CK_j$ is B-NP $\vee$ I-NP$)$
    *for* k $\leftarrow$ 0 to 12 number of seed words
    $Threshold \leftarrow SpacySimilarity(Token_j, seed\ wo$
    *if* $Threshold \geq 60\%$
     $IOBtag_j \leftarrow$ B-FEATURE
    *endif*
   *endfor*
  *endif*
  *endfor*
*endwhile*
*endfor*

checked conditionaly. If both conditons are satisfied then the two token words $TK_j$ and $TK_{j+1}$ are candidate for person entity. The method $Cancatinate(TK_j, TK_{j+1})$ receives current and next tokens and produces a concatination $PersonName$ of these two tokens. The $PersonName$ will be checked in imdbpy software to filter out irrelevant person names as discussed before. If the $PersonName$ exists in imdbpy the $IOBtag_j$ will be labelled B-Person and $IOBtag_{j+1}$ will be labelled I-Person. The abbreviations for variables are TK for Token, CK for chunk and CL for CaseLabel.

### 4.1.4. Semantic Similarity Matrix

The NOUN/NOUN PHRASES will be extracted as potential candidate for aspects; however, the irrelevant aspects will be removed by using SpaCy[4] semantic similarity matrix. Table 6 shows Algorithm III for movie specific aspect pruning. In this process a list of seed words will be matched against each NOUN/NOUNS PHRASE. By calculating the semantic similarity between seed words and potential candidate aspects, the irrelevant aspects will be eliminated. The

**Table 7**

This is an excerpt of relevant detection of movie specific aspects

| Candidate Aspect | Seed Words | Similarity Percentage |
|---|---|---|
| movie | movie | 1.0 |
| film | movie | 0.79121906 |
| star | actor | 0.62115836 |
| horror | movie | 0.61528283 |

semantic similarity percentage below than the specified threshold will be considered irrelevant. In Table 7, the first column shows candidate words and the second column shows aspect seed words and the last column shows similarity percentage.Therefore, in Algorithm III, for every iteration $j$ the current token word's *POS, CaseLabel and Chunk* will be conditionaly checked. If it satisfies, then the candidate movie aspect $Token_j$ will go into a loop of 12 seed words iterations. In any iteration if the semantic similarity of $Token_j$ and $seed\ word_k$ is equal or more than the threshold vlaue which is 60%, then the current token word's $IOBtag_j$ will be labeled as B-FEATURE . The

4 https://spacy.io/usage/vectors-similarity

abbreviations for variables in Algorithm III are TK for Token, CK for chunk, CL for CaseLabel.

The explicit aspects, other than the Person Names and Movie Titles, have also been extracted using POS tag pattern of Uni-gram that is NNP/NN. Table 8 shows the linguistic rule to extract candidate movie specific aspects.

**Table 8**

The output of Algorithm III, depicts the linguistic rule labelled the correct unigram word as movie specific aspect

| Word | POS | Chunk | Case | IOB |
|---|---|---|---|---|
| **scene** | NN | I-NP | LC | B-Feature |
| **film** | NN | I-NP | LC | B-Feature |

From the **Table 8**, if the current word's POS label is NN and its chunking label is B-NP or I-NP and its word case label is LC, then this token word would be tagged with feature Entity. Conversely, this rule also extracts irrelevant movie aspects, for example, Table 9, and shows the irrelevant extraction of movie aspects. The two words brain and wow are not the aspects of any movie but still they have been extracted as movie aspects.

**Table 9**

The output of Algorithm III depicts the linguistic rule labelled the incorrect unigram word as movie specific aspect

| Word | POS | Chunk | Case | IOB |
|---|---|---|---|---|
| **brain** | NN | I-NP | LC | B-Feature |
| **wow** | NN | I-NP | LC | B-Feature |

To remove irrelevant general aspects, a list of seed words and candidate aspects have been provided as input for SpaCy[5] semantic similarity matrix to calculate similarity percentage. Each candidate word have been semantically checked against seed word. If the percetage is more than or equal to 60% threshold value, then the candidate word is a relevant aspect. The seed words are: ("movie", "film", "story", "plot", "character", "scene", "director", "script", "time", "producer", "writer"). The seed words have been discovered from the movie corpus. In this case, the most occurring NOUN has been extracted, Figure 4 shows the empirical representation of the each of these seed word. The most frequent NOUN has been included in the seeds list.

### 4.2. Annotation Phase of Movie Reviews

As shown in the Figure 1, a graphical model of annotation process for aclIMDb dataset. It receives unstructured texts and produces training and testing file. The aclIMDb dataset is based on unstructured data; consequently, the linguistic features have to be extracted. These linguistic features include token sequences, POS tag and chunking information of token sequences. All of these features have been retrieved by using OpenNLP[6] tool. Before extracting these NLP features a review has been shown in Figure 2. The movie review is consisted of words, sentences and paragraph.

**Figure 2**

An Excerpt of an unstructured Movie Review

Story of a man who has unnatural feelings for a pig. Starts out with an opening scene that is a terrific example of absurd comedy. A formal orchestra audience is turned into an insane, violent mob by the crazy chanting of its singers. Unfortunately it stays absurd the WHOLE time with no general narrative eventually making it just too off putting. Even those from the era should be turned off. The cryptic dialogue would make Shakespeare seem easy to a third grader. On a technical level it's better than you might think with some good cinematography by future great Vilmos Zsigmond. Future stars Sally Kirkland and Frederic Forrest can be seen briefly.

After applying tokenization, POS tagger and chunker NLP procedures on the above text, the unstructured text will be converted in to a five column file as shown in Figure 3. Now, the review text is converted into five columns and the first column represents word sequences, second column POS tagger information of the corresponding word, similarly, the third column shows chunking information of the corresponding word.

Next are the orthographic features have been added to the annotation of aclIMDb and the forth column represents orthographic features. These orthographic[6] features identify upper case, lower case and title case of the token words as shown in the Figure 3. Moreover, the possible Labels could be UC, LC, or TC. The fifth column represents IOB (Inside outside Beginning) tagging scheme.

5 https://spacy.io/usage/vectors-similarity

6 https://opennlp.apache.org/

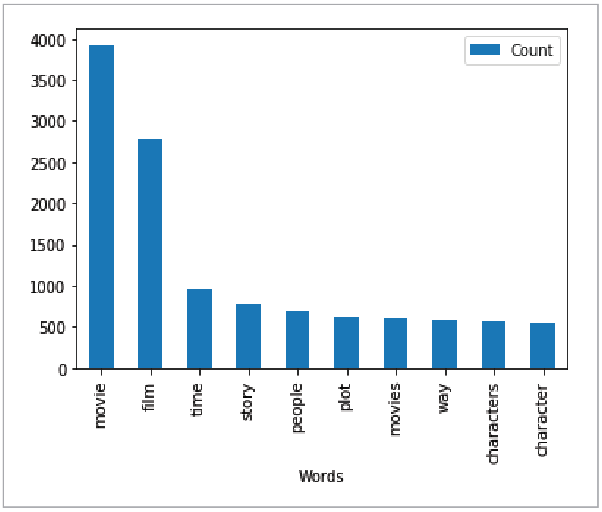
**Figure 3**

An excerpt of an annotated sentence from a movie review

| Word | POS | Chunk | Case | IOB |
|---|---|---|---|---|
| Story | NNP | B-NP | TC | 0 |
| of | IN | B-PP | LC | 0 |
| a | DT | B-NP | LC | 0 |
| man | NN | I-NP | LC | 0 |
| who | WP | B-NP | LC | 0 |
| has | VBZ | B-VP | LC | 0 |
| unnatural | JJ | B-NP | LC | 0 |
| feelings | NNS | I-NP | LC | 0 |
| for | IN | B-PP | LC | 0 |
| a | DT | B-NP | LC | 0 |
| pig | NN | I-NP | LC | 0 |
| . | . | 0 | LC | 0 |

**Figure 4**

Graphical depiction of most occurring words in the corpus of movie reviews




Mainly, four name entities are involved in movie application domain such as PERSON, MOVIE, FEATURE and OPINION. Consequently, the possible IOB tagging would be B-PERSON, I-PERSON, B-MOVIE, I-MOVIE, B-FETURE, I-FEATURE, B-OPINION, I-OPINION and O. Finally, for the opinion words labelling, Table 10 shows the linguistic rules which have been used in the annotation process. If the current POS label is `JJ` and Title Case is `LC`, then the candidate word has been labeled with OPINION. Similarly, if the current chunk is `B-AD` or `B-VP` and Title Case is `LC` then the candidate word has been labeled with OPINION.

**Table 10**

The Linguistic Rule for extracting and labelling the N-gram Opinion Words

| Linguistic Rule | Entity Name |
|---|---|
| *if* $POS_i$ is JJ ^ $CaseLabel_i$ is LC | OPINION |
| *if* $Chunk_i$ is B-AD ˅ B-VP ^ $CaseLabel_i$ is LC | OPINION |

## 5. MAIM Training and Aspect Identification Phase

In this phase, the input will be the Training set of files $T = \{t_1, t_2, t_3, \ldots, t_n\}$ to the classifier. The output will be the identification of entities such as person names, movie titles, explicit aspects and opinion chunks or word. The nature of the dataset that will be used in this research is sequential or in other words unstructured text. The most suitable [17] algorithms for this type of data are *Maximum Entropy model, Hidden Markov Model and Conditional Random Field*. However, among these, CRF is an advanced version of both of these two models, since the aclIMDb dataset is based on sequential data and contains NER problem. Hence, it is the most appropriate classifier for the sequential classification problem. The CRF classifier is capable of identifying sequential data such as human text and biological sequences [31]. Moreover, Chen et al. [14] has identified product aspects and opinion words by using the CRF classifier, whereas Sun et al. [53] has identified name entities and proved that the CRF classifier performs better in detecting NER than the GATE (General Architecture for Text Engineering) and LbjNerTagger. However, Etaiwi et al. [23] proposed a model for the identification of Arabic Names and states that the CRF is very efficient in identifying NER problem.

However, the recent studies [16, 18, 32] concluded that the use of CRF alone will not give a better performance. The amalgamation of CRF and neural networks will provide efficient performance. Consequently, this research will also exploit the capabilities of CRF and Neural Networks (BiLSTM-CRF) [25]. The bidirectional LSTM-CRF uses both features that lie in the past and future (bidirectional pathways) and provides results that are much better than the other machine learning techniqes. Conversely, BiLSTM-CRF performs better than others because it does not reply only on previous and current state. It also exploits the future knowledge and provides efficient performance. In addition, the CRF layer that contains sentence level tagging information will further solve the classification problem.

Figure 5 is a graphical dipection of BiLSTM-CRF. Moreover, Table 11 represents parameter settings for BiLSTM-CRF. After the training phase, the identification phase predicts the label for each token word. Finally, The evaluation of the MAIM has been performed. This assessment has been taken place on the basis of precision, recall and F1 measures.

**Figure 5**

Graphical Representation of five layered BiLSTM-CRF Model

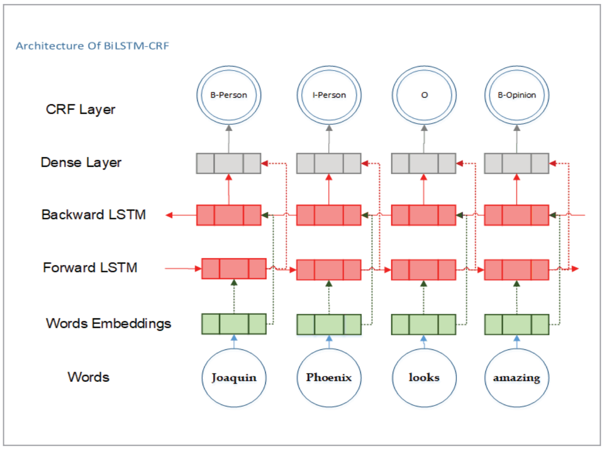



**Table 11**

Parameters and Layers of BiLSTM-CRF, capable of identifying Named Entities

| Layers and Parameters | Input layer | Embedding | Bidirectional | CRF Layer |
|---|---|---|---|---|
| Input Shape | (75,) | N/A | N/A | N/A |
| Output dimension | N/A | 20 | N/A | N/A |
| Input Length | N/A | 75 | N/A | N/A |
| Units | N/A | N/A | 50 | N/A |
| Output Labels | N/A | N/A | N/A | 8 |

## 6. Experimental Results

Large Movie Dataset[7] is the most renowned movie reviews which are a size of 50K, moreover, this dataset has been furthered divided into 25k for training and 25k for testing reviews. Each of them is further divided into 12k positive and 12k negative reviews. The training ratio is 80% and 20% for model training process. CRF, LSTM-CRF and BiLSTM-CRF models have been implemented on aclIMDb dataset. CRF (Conditional Random Field) is an advance version of HMM (Hidden Markov Model) and is capable of modelling overlapping and non-independent features. Moreover, it can be an undirected graph. The CRF classifier is capable of identifying sequential data such as human text and biological sequences [31]. Moreover, Chen et al. [14] has identified product aspects and opinion words by using the CRF classifier, whereas Sun et al. [53] has identified name entities and proved that the CRF classifier performs better in detect NER than the GATE and LbjNerTagger. However, Etaiwi et al. [23] has proposed a model for the identification of Arabic Names and states that the CRF is very efficient in identifying NER problem, since the movie application domain is based on sequential data and contains NER problem. Essentially, CRF is heavily dependent on handcraft features. Moreover, feature generation is a very labouring procedure [25]. In this research four feature set has been used in the training of CRF model, for instance, word token, POS, chunking information and orthographic features.

**Table 12**

A Comparison of MAIM, CRF and LSTM-CRF

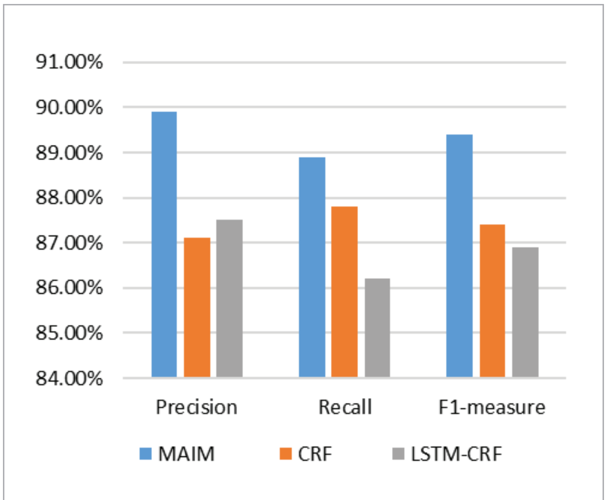
| | Precision | Recall | F1-Measure |
|---|---|---|---|
| **MAIM (Bidirectional LST-CRF without Feature Set)** | 89.9% | 88.9% | 89.4% |
| **CRF(POS, chunk, Case)** | 87.1% | 87.8% | 87.4% |
| **LSTM-CRF (without Feature Set)** | 87.5% | 86.2% | 86.9% |

7 https://ai.stanford.edu/~amaas/data/sentiment/

Neural Network family has been designed for extracting relevant data from sequential data [19]. The major problem with RNN is its biasness towards the most recent input sequences and fails to learn long distance dependencies. In contrast, LSTM is capable of learning long dependencies by using memory cells. However, in this research a hybrid approach has been used LSTM-CRF on movie review dataset. This hybrid approach is capable of using previous features by using LSTM approach and sentence level labelling by using CRF model. CRF uses transition graphs and predicts current tag on the basis of previous and next tag. Therefore, CRF has been used as the output layer.

**Figure 6**

Overall accuracy of MAIM, CRF and LSTM-CRF




For testing and training, the sentences have been divided into tokens and labels in every epoch. The training and testing split is 80% and 20% for training and testing the LSTM-CRF model. The embedding layer is given the parameters such as input length, output dimension, mask_zero; their values are 75, 20 and true respectively. Similarly, LSTM is given the parameters such as units, return_sequences, recurrent_dropdown; their values are 50, true and 0.1 respectively. Then a dense layer has been used as recommended by RNN. Finally, CRF layer has been used as an output layer. The major advantage of LSTM or BiLSTM is their robustness they perform better even if handcrafted features are not being used, consequently, LSTM-CRF or BiLSTM has been trained with only word feature. However, the *MAIM* has been implemented using BiLSTM-CRF on aclIMDb dataset, because BiLSTM-CRF performs better as compared to LSTM-CRF. The training and testing parameters are the same as for LSTM-CRF. Similarly, BiLSTM is also using word feature for the training of the model, moreover, details of BiLSTM-CRF has been provided in section five.

### 6.1. Result Comparison

The overall performance of MAIM is shown in Table 12. The graphical representation achieved accuracies have been shown from Figure 6 to Figure 10. The MAIM (BiLSTM-CRF) performance is better than CRF and LSTM-CRF techniques. In terms of accuracy, the MAIM has 2.8% more precision, 1.1% more recall and 2% more F1-measure than CRF. Similarly, 2.4% more precision, 2.7% more recall and 2.8% more F1-measure as compared to LSTM-CRF technique. In the same way, CRF got 0.4% less precision, 1.6% more recall and 0.5% more F1 measure than LSTM-CRF. The reason CRF is performing better than LSTM-CRF is that it is heavily engineered with features such as POS, chunking and title case information. If these features will be removed, the performance of CRF will be considerably degraded. If entity-wise performance is analyzed, then MAIM (BiLSTM-CRF) still performs better than CRF and LSTM-CRF. For Opinion Entity MAIM (BiLSTM), it got 3% more precision, 2% more recall and F1-meausre is 3% more than CRF. The least recall has been recorded by LSTM in opinion entity, which is 4% less than BiLSTM and 1% less than CRF as shown in Figure 7.

For the MOVIE named entity, MAIM (BiLSTM) performs better in F1-measure and reported 6% and 2% more F1-measure than CRF and LSTM-CRF. However, recall-wise performance has degraded 14%. In contrast, 23% and 1% more precision than CRF and LSTM-CRF has been reported. The reason is that CRF is performing better in recall-wise since it is correctly predicting the total number of actual value, whereas, BiLSTM is performing better while correctly predicting total number of all predictions. In the FEATURE Named Entity CRF model performed 3% more in precision, recall and F1-measure as compared to BiLSTM and LSTM-CRF. The reason is that CRF relying on handcrafted features and if these feature are removed, the performance of CRF will degraded considerably. Finally, for the PERSON named entity the LSTM-CRF performs better 1% than BiLSTM-CRF in precision-wise. However, in recall wise BiLSTM is better 1% and 2% than CRF and LSTM-CRF. Moreover, BiLSTM has 2% and 1% more F1-measure as compared to CRF and LSTM-CRF.

**Figure 7**

Opinion Entity accuracy for MAIM, CRF and LSTM-CRF

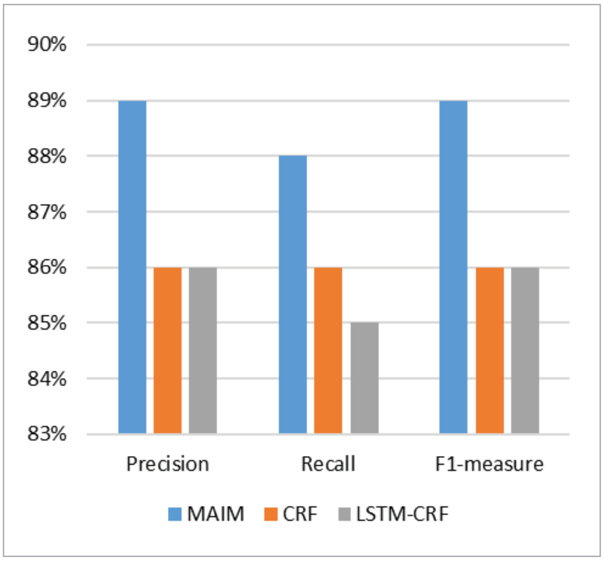



**Figure 8**

Movie Title Entity accuracy of MAIM, CRF and LSTM-CRF

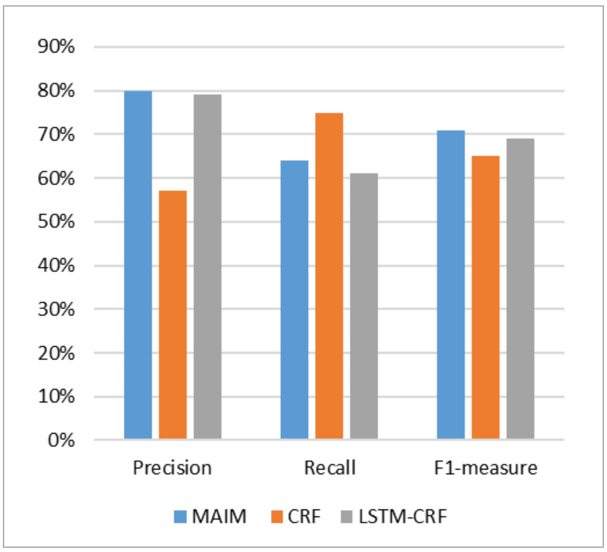



**Figure 9**

Feature Entity accuracy of MAIM, CRF and LSTM-CRF

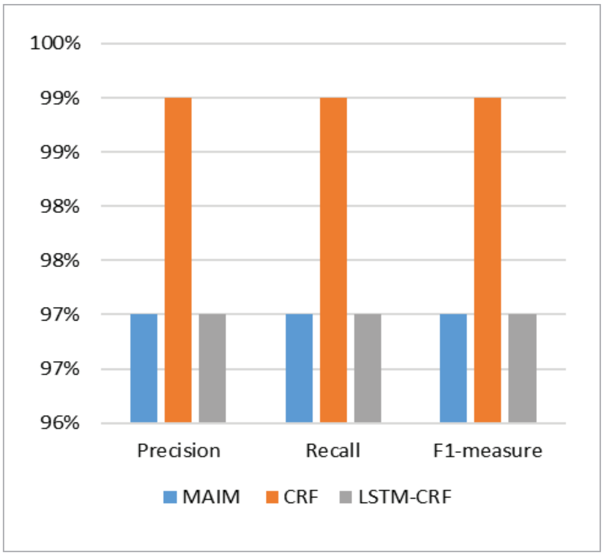
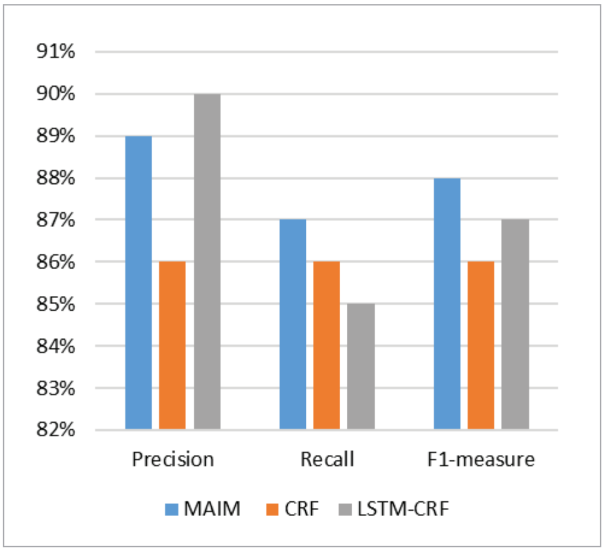



**Figure 10**

Person Entity accuracy of MAIM, CRF and LSTM-CRF




## 7. Discussion

The variants of Recurrent Neural Networks (RNNs) require huge amount of training corpus. Recently, Deep learning based NER models [16, 32] have been validated on the previously annotated datasets. The labelled dataset for movie application is not available [34]. Consequently, the syntactic methods and linguistic rules have been used to annotate Person Names, Movie Titles, Movie specific aspects and opinion words in movie application domain. The irrelevant Named Entities have been eliminated by using imdbpy (internet movie database python) and the irrelevant movie specific aspects have been removed by using spaCy similarity matrix. Even though, this assists to improve annotation process, still many variants of Named Entities cannot be identified. The bigram NOUN PHRASES with title case as orthographic feature have been labelled with Person Name. Nonetheless, other conventions of name writing have been ignored. For instance, "Kathy Bates performed very nice in AHS, however, Bates looked so skinny in this is movie." In this review the writer has written the full name before "however" after that, he has referred the same person with the last name. Since the annotation process accepts only bigram words as Candidate Named Entities, it will ignore the unigram named entity.

In addition, the annotation process is also unable to annotate names longer than two consecutive words, for instance, "William Bradley Pitt" will not be annotated properly and only first two consecutive words will be annotated with Person Entity. Moreover, some writers may only use initials for names, for instance, if the name is "Khalid Khan" the writer may write it like that, "K. Khan". Hence, the annotation process is also not able to annotate that kind of the name version. For movie title entity, the problem is associated with imdbpy. The movie database is too general that cannot distinguish the word "it" as a pronoun or title of a movie. Conversely, there is no way to determine the exact context of the word "it" in a review. Since word "it" also appears as movie name in imdbpy, while annotating movie titles stop words and pronouns have been ignored even if they appear as a movie title in the review.

The Named Entity pruning method is very time consuming. In the first step, a linguistic rule is to extract the candidate NE (Named Entity). After that process, the candidate NE will be match against internet movie database if the candidate NE exists then it will be considered as a valid person name. Unfortunately, if that identified name appears again in the reviews, there is no way to remember that the name is already identified as a valid name. It has to go through a lengthy procedure. Apart from NE, MAIM also classifies N-gram opinion words. These opinion words are sometimes unigram, bigram, trigram, etc. in a sentence of a review. These chunks of opinion words do not just need to be converted to numerical figures such as positive or negative. It is also vital to map implicit aspects to explicit aspects, since these chunks of opinion words may contain movie implicit aspects. Hence, MAIM only identifies explicit movie aspects and opinion words. It does not provide methods for opinion words polarity determination and implicit aspects mapping.

## 8. Conclusion

In this paper, a Movie Aspects Identification Model (MAIM) has been proposed which is capable of identifying movie specific aspects and Named Entities (Person Name and Movie Title). In addition, it also identifies sentiments that have been expressed for movie specific aspects and Named Entities. The significance of this research lies in the introduction of annotation process, which annotates the movie aspects, named entities and sentiment words. MAIM annotation process annotates movie aspects without the discrimination of frequent or infrequent terms. It also provides a movie aspects pruning method in order to remove noisy data from the movie corpus. The renowned BiLSTM-CRF model has been used to identify four main entities PERSON, MOVIE, FEATURE and OPINION. Without any features engineering MAIM performed better than the LSTM-CRF and CRF. It has got 2.8% more precision, 1.1% more recall and 2% F1-measure as compared to CRF and LSTM-CRF.

Although MAIM extracts explicit aspects whether they are named entities or movie specific aspects such as plot, scene or actor, yet it is still incapable of mapping implicit aspects to explicit aspects. The future work will be to map implicit aspects of movie application domain to explicit aspects. For the mapping of implicit aspects, our source will be opinion chunks which have been already identified in the sentence. In this case, an opinion chunk will be mapped to a target explicit aspect on the basis of similarity index or by using conceptNet[8]. Thus, a complete MAIM will be provided which is able to identify both explicit and implicit aspects.

8 https://conceptnet.io/

## References


1. Agerri, R., Rigau, G. Robust Multilingual Named Entity Recognition with Shallow Semi-Supervised Features. Artificial Intelligence, 2016, 23863-23882. https://doi.org/10.24963/ijcai.2017/703
2. Aguilar, G., Maharjan, S., Monroy, A.P.L., Solorio, T. A Multi-Task Approach for Named Entity Recognition in Social Media Data. In Proceedings of the 3rd Workshop on Noisy User-generated Text, 2017.https://doi.org/10.18653/v1/W17-4419
3. Ahmad, T., Doja, M. N. Ranking System for Opinion Mining of Features from Review Documents. International Journal of Computer Science Issues, 2012, 9(4), 440-447.
4. Ahmed, E., Sazzad, M. A. U., Islam, M. T., Azad, M., Islam, S., Ali, M. H. Challenges, Comparative Analysis and a Proposed Methodology to Predict Sentiment from Movie Reviews Using Machine Learning. In 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), 2017. https://doi.org/10.1109/ICBDACI.2017.8070814
5. Al-Smadi, M., Talafha, B., Al-Ayyoub, M., Jararweh, Y. Using Long Short-Term Memory Deep Neural Networks for Aspect-Based Sentiment Analysis of Arabic Reviews. International Journal of Machine Learning and Cybernetics, 2019, 10(8), 2163-2175. https://doi.org/10.1007/s13042-018-0799-4
6. Al-Smadi, M., Qawasmeh, O., Al-Ayyoub, M., Jararweh, Y., Gupta, B. Deep Recurrent Neural Network vs. Support Vector Machine for Aspect-Based Sentiment Analysis of Arabic Hotels‘ Reviews. Journal of Computational Science, 2018, 27386-393. https://doi.org/10.1016/j.jocs.2017.11.006
7. Anand, D., Naorem, D. Semi-Supervised Aspect Based Sentiment Analysis for Movies Using Review Filtering. Procedia Computer Science, 2016, 8486-93. https://doi.org/10.1016/j.procs.2016.04.070
8. Bechara, A. The Role of Emotion in Decision-Making: Evidence from Neurological Patients with Orbitofrontal Damage. Brain and Cognition, 2004, 55(1), 30-40. https://doi.org/10.1016/j.bandc.2003.04.001
9. Bedi, P., Khurana, P. Sentiment Analysis Using Fuzzy-Deep Learning. In Proceedings of ICETIT 2019. 2020, 246-257. https://doi.org/10.1007/978-3-030-30577-2_21
10. Blank, G., Reisdorf, B.C. The Participatory Web: A User Perspective on Web 2.0. Information, Communication & Society, 2012, 15(4), 537-554. https://doi.org/10.1080/1369118X.2012.665935
11. Camilleri, D. Prescott, T. Analysing the Limitations of Deep Learning for Developmental Robotics. In Conference on Biomimetic and Biohybrid Systems, 2017. https://doi.org/10.1007/978-3-319-63537-8_8
12. Chauhan, G. S., Meena, Y. K. DomSent: Domain-Specific Aspect Term Extraction in Aspect-Based Sentiment Analysis. In Smart Systems and IoT: Innovations in Computing, 2020, 103-109. https://doi.org/10.1007/978-981-13-8406-6_11
13. Che, W., Zhao, Y., Guo, H., Su, Z., Liu, T. Sentence Compression for Aspect-Based Sentiment Analysis. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(12), 2111-2124. https://doi.org/10.1109/TASLP.2015.2443982
14. Chen, L., Qi, L., Wang, F. Comparison of Feature-Level Learning Methods for Mining Online Consumer Reviews. Expert Systems with Applications, 2012, 39(10), 9588-9601. https://doi.org/10.1016/j.eswa.2012.02.158
15. Chinsha, T., Joseph, S. A Syntactic Approach for Aspect Based Opinion Mining. In 2015 IEEE International Conference on Semantic Computing (ICSC), 2015.
16. Dernoncourt, F., Lee, J. Y., Szolovits, P. NeuroNER: an Easy-to-Use Program for Named-Entity Recognition Based on Neural Networks. arXiv preprint arXiv:1705.05487, 2017. https://doi.org/10.18653/v1/D17-2017
17. Dietterich, T. G. Machine Learning for Sequential Data: A Review. In Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), 2002.
18. Do, H., Than, K., Larmande, P. Evaluating Named-Entity Recognition Approaches in Plant Molecular Biology. In International Conference on Multi-disciplinary Trends in Artificial Intelligence, 2018. https://doi.org/10.1101/360966
19. Dong, C., Zhang, J., Zong, C., Hattori, M., Di, H. Character-Based LSTM-CRF with Radical-Level Features for Chinese named Entity Recognition. In Natural Language Understanding and Intelligent Applications, 2016, 239-250. https://doi.org/10.1007/978-3-319-50496-4_20
20. Dong, M., Li, Y., Tang, X., Xu, J., Bi, S., Cai, Y. Variable Convolution and Pooling Convolutional Neural Network for Text Sentiment Classification. IEEE Ac-

cess, 2020, 816174-16186. https://doi.org/10.1109/ACCESS.2020.2966726

21. Dragoni, M. A Three-Phase Approach for Exploiting Opinion Mining in Computational Advertising. IEEE Intelligent Systems, 2017, 32(3), 21-27. https://doi.org/10.1109/MIS.2017.46
22. Erickson, R. Are Humans the Most Intelligent Species? Journal of Intelligence, 2014, 2(3), 119-121. https://doi.org/10.3390/jintelligence2030119
23. Etaiwi, W., Awajan, A., Suleiman, D. Statistical Arabic Name Entity Recognition Approaches: A Survey. Procedia Computer Science, 2017, 11357-11364. https://doi.org/10.1016/j.procs.2017.08.288
24. García-Pablos, A., Cuadros, M., Rigau, G. W2VLDA: Almost Unsupervised System for Aspect Based Sentiment Analysis. Expert Systems with Applications, 2018, 91127-137. https://doi.org/10.1016/j.eswa.2017.08.049
25. Huang, Z., Xu, W., Yu, K. Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv preprint arXiv:1508.01991, 2015.
26. Ikram, M. T., Afzal, M. T. Aspect Based Citation Sentiment Analysis Using Linguistic Patterns for Better Comprehension of Scientific Knowledge. Scientometrics, 2019, 119(1), 73-95. https://doi.org/10.1007/s11192-019-03028-9
27. Jibran Mir, A. M., Khatoon, S. Aspect Based Classification Model for Social. Engineering, Technology & Applied Science Research, 2017, 7(6), 2296-2302. https://doi.org/10.48084/etasr.1578
28. Kapočiūtė-Dzikienė, J., Damaševičius, R., Woźniak, M. Sentiment Analysis of Lithuanian Texts Using Traditional and Deep Learning Approaches. Computers, 2019, 8(1), 4. https://doi.org/10.3390/computers8010004
29. Kokina, J., Pachamanova, D., Corbett, A. The Role of Data Visualization and Analytics in Performance Management: Guiding Entrepreneurial Growth Decisions. Journal of Accounting Education, 2017, 3850-62. https://doi.org/10.1016/j.jaccedu.2016.12.005
30. Kumar, R., Garg, S. Aspect-Based Sentiment Analysis Using Deep Learning Convolutional Neural Network. In Information and Communication Technology for Sustainable Development. 2020, 43-52. https://doi.org/10.1007/978-981-13-7166-0_5
31. Lafferty, J., McCallum, A., Pereira, F. C. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, 2001.
32. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C. Neural Architectures for Named Entity Recognition. arXiv preprint arXiv:1603.01360, 2016, https://doi.org/10.18653/v1/N16-1030
33. Lerner, J. S., Li, Y., Valdesolo, P., Kassam, K. S. Emotion and Decision Making. Annual Review of Psychology, 2015, 66799-667823.
34. Li, J., Sun, A., Han, J., Li, C. A Survey on Deep Learning for Named Entity Recognition. IEEE Transactions on Knowledge and Data Engineering, 2020. https://doi.org/10.1109/TKDE.2020.2981314
35. Li, S., Zhou, L., Li, Y. Improving Aspect Extraction by Augmenting a Frequency-Based Method with Web-Based Similarity Measures. Information Processing & Management, 2014, 51(1), 58-67. https://doi.org/10.1016/j.ipm.2014.08.005
36. Limsopatham, N., Collier, N. Learning Orthographic Features in Bi-Directional LSTM for Biomedical Named Entity Recognition. In Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016), 2016.
37. Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., Potts, C. Learning Word Vectors for Sentiment Analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, 2011.
38. Manek, A. S., Shenoy, P. D., Mohan, M. C., Venugopal, K. Aspect Term Extraction for Sentiment Analysis in Large Movie Reviews Using Gini Index Feature Selection Method and SVM Classifier. World Wide Web, 2017, 20(2), 135-154. https://doi.org/10.1007/s11280-015-0381-x
39. Marrese-Taylor, E., VelÃisquez, J. D., Bravo-Marquez, F. A Novel Deterministic Approach for Aspect-Based Opinion Mining in Tourism Products Reviews. Expert Systems with Applications, 2014, 41(17), 7764-7775. https://doi.org/10.1016/j.eswa.2014.05.045
40. Mir, J., Usman, M. An Effective Model for aspect Based Opinion Mining for Social Reviews. In 2015 Tenth International Conference on Digital Information Management (ICDIM), 2015. https://doi.org/10.1109/ICDIM.2015.7381851
41. Panda, M. Developing an Efficient Text Pre-Processing Method with Sparse Generative Naive Bayes for Text Mining. International Journal of Modern Education and Computer Science, 2018, 10(9), 11. https://doi.org/10.5815/ijmecs.2018.09.02
42. Parkhe, V., Biswas, B. Sentiment Analysis of Movie Reviews: Finding Most Important Movie Aspects Using Driving Factors. Soft Computing, 2016, 20(9), 3373-3379. https://doi.org/10.1007/s00500-015-1779-1

43. Peng, H., Ma, Y., Li, Y., Cambria, E. Learning Multi-Grained Aspect Target Sequence for Chinese Sentiment Analysis. Knowledge-Based Systems, 2018, 148167-148176. https://doi.org/10.1016/j.knosys.2018.02.034

44. Pham, D.-H., Le, A.-C. Learning Multiple Layers of Knowledge Representation for Aspect Based Sentiment Analysis. Data & Knowledge Engineering, 2018, 11426-11439. https://doi.org/10.1016/j.datak.2017.06.001

45. Poria, S., Cambria, E., Gelbukh, A. Aspect Extraction for Opinion Mining with a Deep Convolutional Neural Network. Knowledge-Based Systems, 2016, 10842-49. https://doi.org/10.1016/j.knosys.2016.06.009

46. Rana, T. A., Cheah, Y.-N. A Two-Fold Rule-Based Model for Aspect Extraction. Expert Systems with Applications, 2017, 89273-89285. https://doi.org/10.1016/j.eswa.2017.07.047

47. Salas-Zárate, M. d. P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodriguez-Garcia, M. A., Valencia-Garcia, R. Sentiment Analysis on Tweets About Diabetes: An Aspect-Level Approach. Computational and Mathematical Methods in Medicine, 2017. https://doi.org/10.1155/2017/5140631

48. Sang, E. F., De Meulder, F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. arXiv preprint cs/0306050, 2003.

49. Schouten, K., Frasincar, F. Survey on Aspect-Level Sentiment Analysis. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(3), 813-830. https://doi.org/10.1109/TKDE.2015.2485209

50. Schouten, K., Van Der Weijde, O., Frasincar, F., Dekker, R. Supervised and Unsupervised Aspect Category Detection for Sentiment Analysis with Co-occurrence Data. IEEE Transactions on Cybernetics, 2017, 48(4), 1263-1275. https://doi.org/10.1109/TCYB.2017.2688801

51. Şeker, G. A., Eryiğit, G. Extending a CRF-based Named Entity Recognition Model for Turkish Well Formed Text and User Generated Content 1. Semantic Web, 2017, 8(5), 625-642. https://doi.org/10.3233/SW-170253

52. Shams, M., Baraani-Dastjerdi, A. Enriched LDA (ELDA): Combination of Latent Dirichlet Allocation with Word Co-occurrence Analysis for Aspect Extraction. Expert Systems with Applications, 2017, 80136-80146. https://doi.org/10.1016/j.eswa.2017.02.038

53. Sun, B. Named Entity Recognition: Evaluation of Existing Systems. Institutt for datateknikk og informasjonsvitenskap, 2010.

54. Tripathy, A., A. Anand, and S.K. Rath, Document-level sentiment classification using hybrid machine learning approach. Knowledge and Information Systems, 2017, 53(3), 805-831. https://doi.org/10.1007/s10115-017-1055-z

55. Wiebe, J. M., Bruce, R. F., O‘Hara, T. P. Development and Use of a Gold-Standard Data Set for Subjectivity Classifications. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, 1999. https://doi.org/10.3115/1034678.1034721

56. Yadav, V., Bethard, S. A Survey on Recent Advances in Named Entity Recognition from Deep Learning Models. In Proceedings of the 27th International Conference on Computational Linguistics, 2018.

57. Yin, W., Kann, K., Yu, M., Schütze, H. Comparative Study of CNN and RNN for Natural Language Processing. arXiv preprint arXiv:1702.01923, 2017.

58. Zainuddin, N., Selamat, A., Ibrahim, R. Hybrid Sentiment Classification on Twitter Aspect-Based Sentiment Analysis. Applied Intelligence, 2018, 1-15. https://doi.org/10.1007/s10489-017-1098-6

59. Zeng, L., Li, F. A Classification-Based Approach for Implicit Feature Identification. In Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. 2013, 190-202. https://doi.org/10.1007/978-3-642-41491-6_18

60. Zhao, Y., Qin, B., Liu, T. Creating a Fine-Grained Corpus for Chinese Sentiment Analysis. IEEE Intelligent Systems, 2015, 30(1), 36-43. https://doi.org/10.1109/MIS.2014.33

61. Zhu, J., Zhang, C., Ma, M. Y. Multi-Aspect Rating Inference with Aspect-Based Segmentation. IEEE Transactions on Affective Computing, 2012, 3(4), 469-481. https://doi.org/10.1109/T-AFFC.2012.18