The Journal of Supercomputing (2019) 75:2584–2602
https://doi.org/10.1007/s11227-018-2664-3

A segregational approach for determining aspect sentiments in social media analysis

Asif Nawaz[1] · **Sohail Asghar**[2] · **Syed Husnain Abbas Naqvi**[1]

Published online: 26 October 2018

Abstract

Aspect-based sentiment analysis is an emerging field of research that evaluates people's views, ideas or sentiments. It is a subtask of sentiment analysis that is used to identify text sentiment orientation towards different aspects of a mobile phone such as camera and screen resolution. During the last decade, research community focused on identifying and extracting aspects like the most common methods used for aspect extraction to identify the main features of an entity only. These techniques are corpus or lexicon based and domain specific. Some approaches for aspect extraction are based on term frequency and inverse document frequency. Such approaches are quite good if aspects are associated with predefined categories and may fail if low-frequency aspects are concerned. The heuristic-based approaches are better than frequency and lexicon-based approaches in terms of accuracy, but due to the different combinations of features, they consumed time. The researchers have already implemented machine learning techniques to analyse sentiments present in the given document. But, execution time for these techniques increases due to the increasing aspects in a set of data. Also, irrelevant and redundant aspects participate in determining the sentiment of the given document, thereby varying the accuracy of the algorithm. In this research, we present a segregational approach for aspect identification that is based on aspect and opinion words disentangling and aspect refinement using concept similarity. To obtain better accuracy, we also built a set of part of speech tagger and integrated it with our proposed technique. The experimental analysis reveals that our proposed technique outperforms the existing counterparts.

Keywords Sentiment analysis · Aspects · Aspects extraction · Conceptual similarity

✉ Asif Nawaz
asif.nawaz@uaar.edu.pk

Extended author information available on the last page of the article

1 Introduction

In today's internet age, most people use social media as a source of communication, information sharing and entertainment. They join social media forums to share their views in a form of comments, likes or reviews [1, 2]. Social networks and microblogging sites like Twitter, Facebook and Flicker get more popularity as compared to other print or electronic media because of its wide usage and easy processing [3].

To know company strategies, political movements, social events and product preferences from public views are challenging task. In the past few years, the research community focused on extracting ideas from reviews, tweets and blogs. Many researchers consider this process in two steps: where the first step is about the retrieval of the relevant document topic or reviews by giving user query, and then to re-rank document/topic [4]. The second step which is Opinion Mining (OM) typically composed of two additional subtasks, the first task is to find opinions which extract opinionated posts and reviews about a particular entity by putting relevant user query, and the second task is polarity which recognizes that how a particular review is close to a topic which may be negative, positive or neutral.

Many existing approaches for the sentiment analysis like [5–8] attempted to detect the sentiment at sentence, phrase or paragraph without considering the entities (e.g. laptop or mobile) and their aspects, e.g. memory, battery expressed in context. However, in such approaches when overall sentiments are considered, sometimes there exist chances of failure to capture the sentiments over different aspects on which a particular entity reviews [9]. This problem usually degrades the accuracy because of lacking an efficient aspect identification technique [10].

Abstract-based sentiment analysis (ABSA) is the subtask of sentiment analysis that interprets text sentiments orientation with respect to different aspects. In a document, aspect is a concept on which the reviewer expresses his opinion. For example, consider a review comments like "the optics of this camera is very good and battery life is excellent." From such comments, we can clearly conclude that the review about camera quality is positive; however, the reviewer likes its optics and battery life. The aspect extraction task which is usually known as information extraction detects most optimal aspects from text reviews. Aspects can be categorized into two different types; explicit and implicit aspects. Explicit aspects correspond to specific terms that explicitly appear in the document, whereas implicit aspect is not specified explicitly in the document.

In this research, we propose a novel technique for aspect identification based on aspect and opinion words disentangling and associative similarity along with part of speech tagger. We apply the proposed technique to each word in an opinionated sentence to consider it either that is an aspect or non-aspect word. We also integrate part of speech tagger with the proposed technique to obtain significantly better accuracy than existing approaches.

1.1 Research contributions

The silent features of our research work are narrated as follows:

1. This research tries to attempt aspect identification based on normalized Google distance and concept net methods that can be further used to improve the accuracy and efficiency of sentiment analysis process.
2. By using aspect identification method, the proposed method supersedes the existing state-of-the-art approaches in achieving high accuracy of almost 99%.
3. The aspect reduction process reduces the dimensionality of aspects and thus enhances the performance of different decision making processes.

Rest of the paper is organized as follows: Sect. 2 describes related work. The proposed model for aspect identification is described in Sect. 3. Section 4 provides the detailed description of experiments and evaluation. Discussion on various experiments is described in Sect. 5. Section 6 elaborates conclusion and future work.

2 Related work

There exists many techniques that are used for sentiment analysis in various domains. Most of these mainly focus on discovering sentiments at sentence [11] or at a document [12] level. When these approaches concentrate on sentence level or document level, as a result, they cannot satisfy user's expectations. Mostly, they need perfect information from a sentence or review in terms of specific feature or aspect related to a product. For example, someone only interested in specific features of a mobile product such as "camera result", "screen display" or "design". For always, users see a specific product from different perspectives. So, it is very necessary to discover sentiments in terms of aspects. This mechanism is called as "aspect-based sentiment analysis".

In [11] aspect extraction from opinion was first time studied by stating the difference between implicit and explicit aspects. The implicit aspect is saturated instead of explicit one, so in this paper we mainly focus explicit aspects. Further, in [12] and [13], the above-mentioned approach was improved. The existing work uses the supervised method and point-wise mutual information to judge a product feature by calculating scores of mutual information. But, there have been some gaps in supervised method which has been covered by topic modelling.

Topic modelling has been widely used as a basis to perform aspect extraction and grouping. Existing work shows two models that are PLSA [14] and LDA [15]. PLSA and LDA estimate the semantic topic distribution of any document by suggesting the underlying word "topic" in mid of "document" and "word". Here each document is the copy of random latent topics and each topic is the group of multiple words [16]. A lot of work is done with these models which shows these models are helpful in extracting global (i.e. category) and also local aspects, i.e. camera [17]. These models also provide better results in key phrases extraction and in judging sentiments. Ref. [18] uses the part of speech taggers with maximum entropy that identifies aspects and sentiments.

The authors in [19] implement the three sentiment analysis algorithms for identifying the sentiments (positive or negative) from reviews. Experimental results are then compared with the numerical ratings of hotels. The dataset of one million reviews with a numerical rating is collected from Tripadvisor. Results show that predicted rating from sentiment analysis algorithms is very close to actual ratings of the hotel.

The sentiment analysis techniques are further studied for web-based patient reviews as proposed in [20]. The authors describe how sentiment analysis is useful for patients, doctors as well as a healthcare manager. The author identifies some hurdles which are primarily related to the collection of patient reviews which is very hard to judge exact patient diseases from review comments. Extracting knowledge from reviews using data mining techniques got popularity in the literature and a lot of work has been done on judging the sentiments from a text (i.e. reviews, blogs, tweets), but in [21] the author proposed the open-ended qualitative LibQUAL technique for classifying the comments in positive and negative classes. They have applied the proposed technique on Canadian mid-sized academic research library comments. The achieved results are compared with machine learning algorithms.

The author in [22] proposed a novel algorithm for aspect extraction based on concept parser to get semantic and synthetic aspects in natural language text. Poria [23] explored common sense knowledge and dependency relations to get implicit and explicit aspects. In [24], the author elaborates a dependency tree-based method to generate the most optimal aspects. Zheng in [25] proposed a two-step model that extracts opinion pattern from text reviews. Another approach for aspect extraction was proposed in [26] that use different dependency combinations to extract aspects. In [27] a manual-based dependency rule was analysed in order to link opinion words with target process. Samha in [28] deployed extraction rule to get the best dependency combination for aspect extraction.

From the above discussion, we conclude that most techniques frequently use extraction based on nouns phrases and nouns. Such techniques perform well when many different aspects are fully associated with some categories of terms or words (e.g. nouns), but they either fail when used for terms having low frequency as aspects. Existing literature on sentiment analysis may not focus on the systematic way for feature selection. There are many past approaches that are based on the heuristic methods for the selection of best features or candidates of a classifier for making an ensemble [29]. This process is time-consuming because it is difficult to try different groups of features to finally fix a model. There is another big issue of a specific domain, in which a system is developed which is usually failed when applied on some other domain. It shows that a set of features having good performance for a specific domain cannot perform same for the other domain. Therefore, it is very important to choose aspects carefully. For the selection of aspects in sentiment analysis, the effective use of information gain and Z-score is reported in [30]. An efficient technique of aspect selection for sentiment analysis is presented in [31], and this technique is based on the frequency of documents. However, as compared to information gain-based approach to feature selection its accuracy is low. The notations used in this paper are described in Table 1.

Table 1 Nomenclature

Notation	Description	Notation	Description
ABSA	Aspect-based sentiment analysis	VB	Verb
TF-IDF	Term frequency and inverse document frequency	NN	Noun
OM	Opinion mining	NP	Noun phrase
PLSA	Probabilistic latent sentiment analysis	NNP	Proper noun
LDA	Latent Dirichlet allocation	JJ	Adjective
LibQUAL	Library used for quality services	DT	Article
NGD	Normalized Google distance	WWW	World Wide Web
REST API	Representation state transfer application programming interface	UMBEL	Upper mapping and binding exchange layer
OAuth	Open authorization	GWAP	Game with a purpose
JSON	JavaScript object notation	AR	Aspect reduction
DVD	Digital versatile disc	PA	Proposed approach
MP3	MPEG Layer-3	W2VLDA	A novel method used for ABSA
NLP	Natural language processing	LABSA	Level aspect-based sentiment analysis
POS	Part of speech	PT	Proposed technique

3 Proposed model of aspect identification using NGD and ConceptNet

Given a large corpus of reviews or tweets about any product or entity that identify its refined aspect is a fine-grained task in sentiment analysis. This task is challenging especially when there is no labelled data which can be simulated in a given domain. To address this problem, we propose a general two-stage approach. Stage one extracts and groups the target-related words for a given target using normalized Google distance. This is relatively easy as we can apply an existing semantics-based learning technique. The second stage reduces the aspects by eliminating redundant and irrelevant aspects thereby varying the accuracy of the algorithm. Besides this, the proposed research method revolves around three different contexts, the research process, the data set that will be used for evaluation, and the most important is performance measure indicators that will be utilized.

3.1 The research process

In order to get more successful method, we make sure that the proposed approach performance should be equal or better than today's accepted solutions for aspect identification problems. In the domain of sentiment analysis especially in aspect iden-

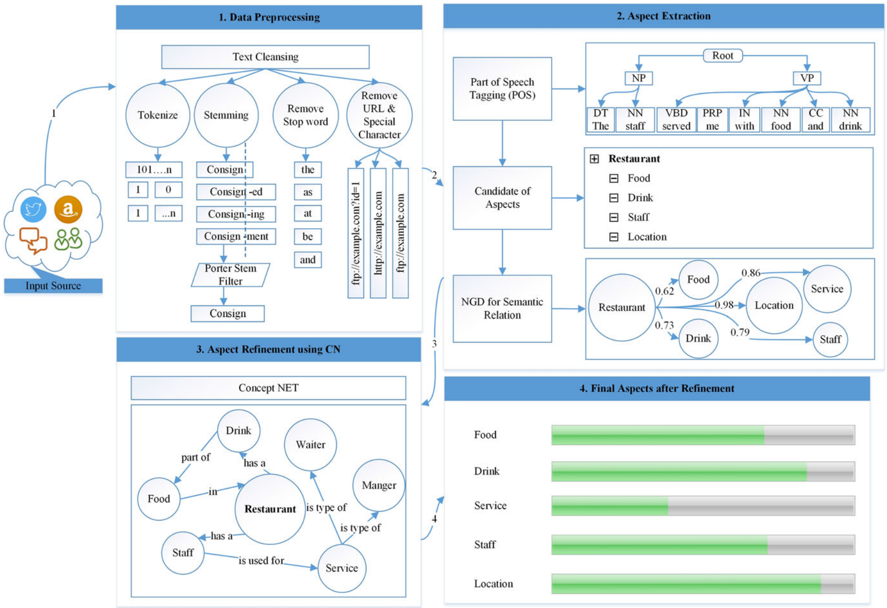

Fig. 1 Proposed model of aspect identification using NGD and ConceptNet

tification, the proposed approach is compared against machine learning approaches that have much better acceptance and credibility. The proposed model is shown in Fig. 1.

3.2 Data set used

The common question that is to be answered is what dataset is used? Below subsections briefly answer this question.

3.2.1 Twitter data set

On the basis of terms and conditions for the purpose of utilization and privacy preservation, most of the twitter datasets were obtained from twitter by public request. However, some of the twitter datasets still exist physically. We choose one of them that were extracted from the Twitter REST API; with the help of REST API, we access twitter data programmatically. With the help of this API, we can read author profiles and followers data. This API uses Open Authorization (OAuth) to identify users and Twitter applications, and it returns a response in JSON form JavaScript object notation [32]. This data set provides a corpus of almost eighty thousand tweets. Each tweet is individually organized in a single row in .xls format. We have randomly chosen almost ten thousand tweets for aspect identification in our experiments. This dataset required much more preprocessing because there were numerous errors, emotional

Table 2 Dataset distribution

Dataset name	Number of review	Vocabulary size	Words in lexicon
Tweet DS1	128,234	28,345	2214
Canon	22,213	19,345	1560
Nikon	11,003	4700	1123
Cell phone	3700	1526	932
DVD/MP3	2500	900	350

icons, mistype statements and strange characters that needed purification as shown in Table 2.

3.2.2 Reviews data set

META-SHARE, the open language resource exchange facility [33] provides the data sets that were utilized in sentiment analysis and experimental evaluations. This data set was further divided into four items; Canon, Nikon, Cell phone, DVD and MP3. The description of the above data set is given in Table 2.

3.3 Preprocessing

Algorithm 1 expresses the details of preprocessing phase. It receives a text as input. The text is fragmented into sentences. With the use of Stanford CoreNLP [34], the POS tagging [35] the stemming process is completed. Sentences such as “this is not a good phone” use words with positive polarity, while the negation word of NOT changes the polarity of a sentence. Because unigram features do not model relations between words in the text, the effect of the negative word should be reflected in the unigram features before their extraction.

Algorithm 1: Preprocessing

```
Input: A Set of Reviews (Di)
Output: All Preprocessed Words (W^all)
Initialize W^all ←φ
    for all D ϵDi
        String S ←vectorization(Di)
        S' ←Splitter(S)
    end for
    for all Si ϵ S
        S" ←POS tagger (Si)
        S"' ←Stemming (S")
    end for
return W^all
end
```

3.4 Aspect extraction using POS tagger

Review/tweet about any product may contain different types of aspects which can be commented individually. In this phase, each sentence of a particular review can be considered as a bag of words. All aspects from this bag are identified using the Modified POS tagger. A POS tagging tool finds out part of speech in a sentence along with grammatical relationships present in a respected sentence. Each word in a sentence is tagged as verb (VB), noun (NN), noun phrase (NNP), proper noun (NNP), adjectives (JJ) and article (DT), etc. However, nouns and proper nouns are usually identified as candidate aspects using POS tagging as shown in Table 3.

3.5 Aspect co-occurrence using NGD

The context of phrases and words varies as per their use in daily life as compared to the semantics to some other phrases and words. For example, in terms of computer "society" can be considered as "database" and "use" is considered as a way that is used for the database. For a particular query, we almost use the World Wide Web (WWW) as a database and use search engine Google. This concept is then further applied to construct a technique that automatically extracts all those pages pertaining to a particular word association using Google page count as shown in Algorithm 2. Page count association measured between different words using normalized Google distance (NGD) was proposed in [36]. Finding association between two different terms using NGD does not require any background knowledge or any particular analysis of problem domain. Instead, it automatically analyses all features through Google search using World Wide Web. The Normalized Google Distance (NGD) is defined in Eq 1.

$$\mathrm{NGD}(\alpha, \beta) = \frac{f(\alpha, \beta) - \min(f(\alpha), f(\beta))}{\max(f(\alpha), f(\beta))} \tag{1}$$

and

$$\mathrm{NGD}(\alpha, \beta) = \frac{\max\{\log f(\alpha), \log f(\beta)\} - \log(\alpha, \beta)}{\log N - \min\{\log f(\alpha), \log f(\beta)\}} \tag{2}$$

where $f(\alpha)$ represents the number of pages that contain term α and $f(\alpha, \beta)$ indicates association of both terms α and β reported by Google. In Eq. 2, N shows number of pages returns from Google. It seems that by decreasing *N,* we increase NGD. Algorithm 2 shows word similarity using NGD technique. In this experiment some main properties of NGD that were applied are as follows:

1. The approximate value of the NGD lies between 0 and ∞, may be sometimes little bit negative if the Google search count irrelevant score like

$$f(\alpha, \beta) > \max\{f(\alpha), f(\beta)\} \tag{3}$$

   (a) in all situations of αandβ if the frequency $f(\alpha) = f(\beta) = f(\alpha, \beta) > 0$, then $\mathrm{NGD}(\alpha, \beta) = 0$.

Table 3 Aspect term extraction using POS tagging

Aspects							
Relationship	Intensive	Civilian	Big	Data	Excellence	Awards	School
Project	Provinces	Link	Joy	Garden	Street	Peace	Prize
Winner	Delegation	Influx	Youth	Side	Cricket	Nation	Heroes
Shortlist	Laboratories	Attack	Mission	Air	Water	Milk	Shame
Returns	Chairman	Fast	Wedding	Fans	Self	Habit	Nation
Years	Enterprise	Neck	Today	Female	Pacer	Entry	Airport
Experience	Pilgrimage	Chamber	Scandal	Rails	Incident	Hours	Parts
Sanctity	Representation	Staff	Vote	History	Media	Evening	Talk
Response	Politicians	Water	Milk	Group	Mission	Shows	Air
People	Enemy	Country	War	Global	Hand	Site	Party
Face	Announcement	Time	Attack	Trip	Hospital	Increase	Self
Retailers	Theatres	Pride	Former	Economic	Problems	Car	Point
Convener	Competitive	Catheter	Exports	Prices	Retail	Family	Guard
Advisor	Appearance	Member	Strategy	Education	Partner	Squad	World
Marketing	Congratulation	Face	Analytics	Run	Agenda	Idea	Flight
Commerce	Government	Strategies	Mud	Insights	Reduce	Delays	Fries
Time	Differences	Leak	Fears	Film	Products	Woods	Trees
Campaigns	Actionable	Nothings	Build				

Table 4 Refine results using NGD

	Data	Excellence	Awards	Project	Staff	Link
Data	1	– 0.011	0.023	0.005	– 0.027	0.174
Excellence	– 0.011	1	0.485	0.037	0.051	0.043
Awards	0.023	0.485	1	0.074	– 0.026	0.05
Project	0.005	0.037	0.074	1	0.065	0.057
Staff	– 0.027	0.051	– 0.026	0.065	1	– 0.008
Link	0.174	0.043	0.05	0.057	– 0.008	1

(b) if the frequency $f(\alpha) = 0$ then for every search term β, we have $\mathrm{NGD}(\alpha, \beta) = \infty/\infty$.

2. The value of NGD is almost non-negative and $\mathrm{NGD}(\alpha, \alpha) = 0$ for every α. For every pair of αandβ, we have $\mathrm{NGD}(\alpha, \beta) = \mathrm{NGD}(\beta, \alpha)$. For example, let x denote the number of web pages containing one or more occurrence of α. For example, choose $\alpha \neq \beta$ with $x = y$ then $f(\alpha) = f(\beta) = f(\alpha, \beta)$ and $\mathrm{NGD}(\alpha, \beta) = 0$. The NGD does not satisfy triangular inequality:

$$\mathrm{NGD}(\alpha, \beta) \leq \mathrm{NGD}(\alpha, \eta) + \mathrm{NGD}(\beta, \eta) \text{for all } \alpha, \beta, \eta \quad (4)$$

where α, β, η represents terms.

Algorithm 2 Compute Word Similarity using NGD

```
1.  Input  :  All Pre-processed words (W^all)
2.  Output:  List of specific datbase words WSD^NGD
3.     Initialize  WSD^NGD ←φ
4.       Cache C←φ.
5.          for all w_i, w_j ϵ W^all
6.             if (w_i, w_j ∉ C )
7.                 S(w_i, w_j) ← Compute _NGD (w_i, w_j)
8.                 C = C U S( w_i, w_j)
9.                 WSD^NGD  ← C
10.            end if
11.         end for
12.    return  WSD^NGD
13.    end
```

Some of the results that were obtained using NGD are shown in Table 4.

The integration of NGD along with part of speech tagger purifies the aspects. From experimental evaluation, we clearly observe promising results. It increases the optimality of aspect extraction process up to 5–6% from part of speech tagger. The obtained results are captured in Table 5 and graphically presented in Fig. 2.

Table 5 Comparison between POS Tagger and NGD

#	Dataset	Reviews	Part of speech tagger	Normalized Google distance
1	Tweets	10,000	84.26	90.15
2	Canon	45	83.41	89.85
3	Nikon	34	86.12	89.54
4	DVD	41	84.27	88.89
5	MP3	95	85.10	90.83
6	Cell phone	99	83.49	88.52

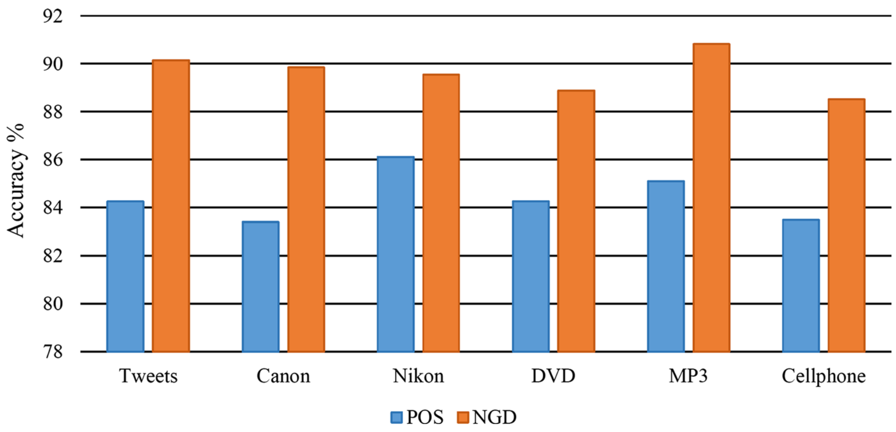

Fig. 2 Comparison between POS and NGD

3.6 Aspect reduction using concept net search

The research community focused on implementing machine learning techniques to analyse sentiments present in the given document. But one of the limitations is that the execution time for these techniques increases due to the increase volume in aspect set of data. Also, irrelevant and redundant aspects participate in determining the sentiment of the given document, thereby varying the accuracy of the algorithm. The main goal of this phase is to decrease the dimensionality of the aspect space and thus computational cost.

ConceptNet is the large network designed for understanding the semantics of words in 1999. Some steps related to the deployment of our conceptual model are given below.

1. All versions of ConceptNet contain relational knowledge of English language, and its sister project contains knowledge of other popular languages.
2. ConceptNet uses the subset of DBpedia [37]; it extracts knowledge from Wikipedia articles. It also uses the multi-lingual dictionary named as Wiktionary. This dictionary provides information about more than hundred languages.
3. WordNet multi-lingual dictionary is also used for extended knowledge.

Table 6 Aspect reduction

No.	Aspect
1	Data
2	Project
	.
	.
	.
	.
	.
	.
.	Staff
.	Link
.	Peace
.	Farmer
.	Film

4. For Semantic Web representation, ConceptNet uses Upper Mapping and Binding Exchange Layer (UMBEL) [38].
5. Some knowledge about peoples sentiments are taken from "Games with Purpose" library. This game was developed by Japanese for project GWAP [39].

A refine model of ConceptNet that was used in our proposed approach is shown in Fig. 3 in the form of tree. Results obtained after applying ConceptNet for aspect refinement are shown in Table 6.

4 Experimental results

In this section various experiments are performed on different datasets in order to evaluate the performance of proposed approach. The detailed description of each experiment is given below.

The first experiment is for aspect reduction (AR) from the large set of aspects. Aspect reduction reduces the set of aspects by removing usual and ordinary aspects that enhances the process of sentiment analysis up to 30–40%. Comparison of the proposed AR technique with hybrid approach [40] is shown in Table 7, which is sketched in Fig. 4.

Various experiments are performed on Twitter and Reviews datasets for validating the results. As shown in Table 8, our proposed approach yields accuracy, precision and recall (Fig. 5).

Table 9 shows the comparison in terms of accuracy with two latest techniques that are W2VLDA [41] and LABSA [42]. "W2VLDA" is a system for aspect-based sentiment analysis. This is a technique uses the topic modelling technique with other approaches to model the topic names and assign the sentiment to classes. A second one is the "Level Aspect-Based Sentiment Analysis Using Ontology", which uses a new method of ontology to perform sentiment analysis. A graphical representation is shown in Fig. 6.

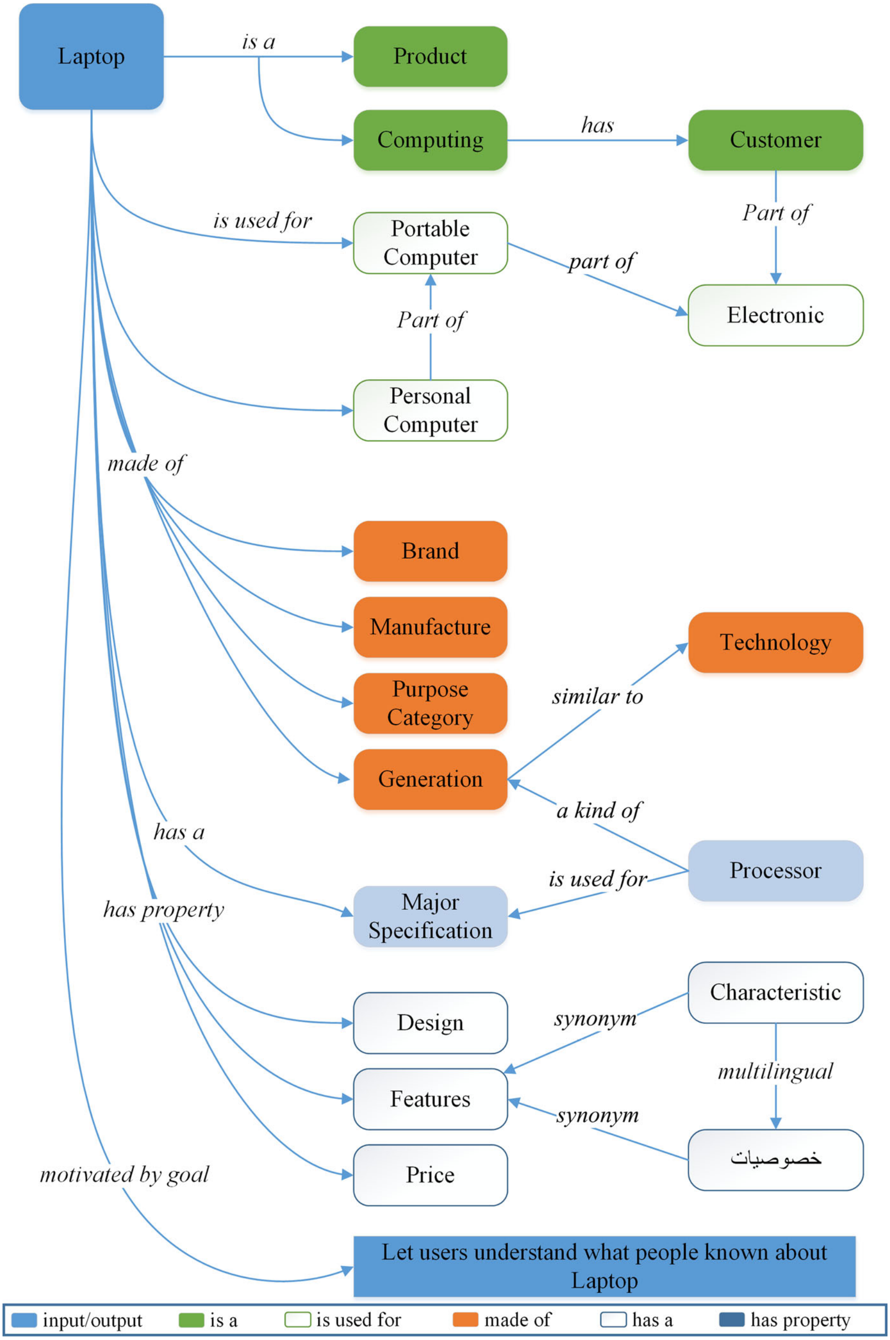

Fig. 3 Tree of different aspects using ConceptNet

Table 7 Comparison of aspect reduction techniques

#	Dataset	Reviews	Proposed approach (PA)	Ontology based approach (UA) [42]
1	Tweets	10,000	71.33	65.15
2	Canon	45	79.51	72.85
3	Nikon	34	76.28	70.45
4	DVD	41	81.72	73.98
5	MP3	95	79.10	72.38
6	Cell phone	99	77.94	71.25

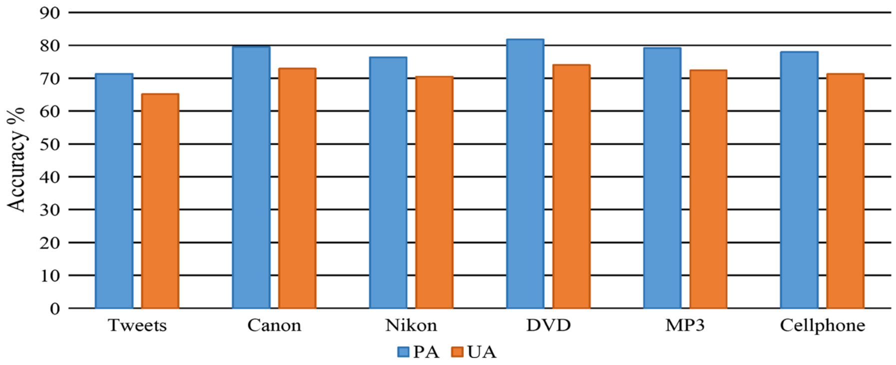

Fig. 4 Comparison of aspect reduction techniques

Table 8 Experimental results in terms of accuracy, precision and recall

#	Domain	Dataset	Precision	Recall	Accuracy
1	Twitter	Tweets	86.05	85.24	87.45
2		Canon	92.07	90.24	93.05
3		Nikon	89.36	88.17	91.56
4	Reviews	DVD	91.82	91.07	91.74
5		MP3	92.74	90.25	92.36
6		Cell phone	90.45	89.16	90.85

From the above debate, we observed that the results obtained are very encouraging. The proposed technique improves the results in both Twitter and product review datasets. Table 5 clearly shows that our proposed technique performs well as compared to the other state-of-the-art techniques and improves the results up to 4–8%. It is only possible by aspect reduction phase as it removes the unwanted and redundant aspects.

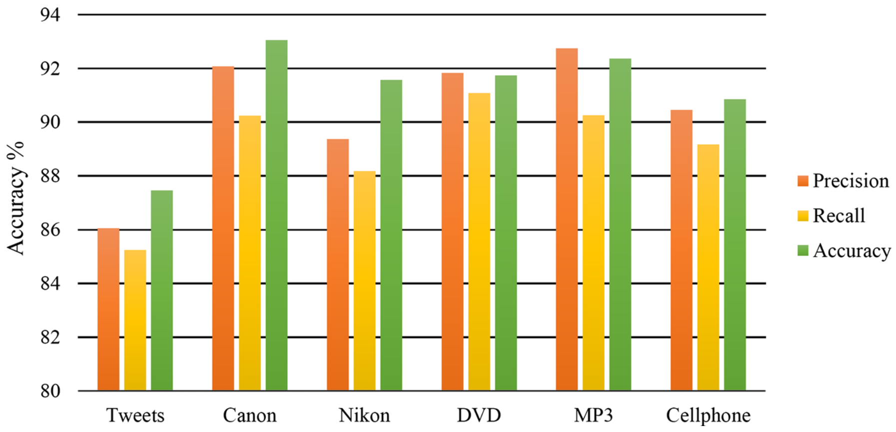

Fig. 5 Experimental results in terms of Precision, Recall and Accuracy

Table 9 Comparison results of W2VLDA, LABSA and P.T

#	Dataset	W2VLDA	LABSA	Proposed technique (P.T)
1	Tweets	81.84	78.12	87.45
2	Canon	86.24	84.24	93.05
3	Nikon	87.74	81.95	91.56
4	DVD	85.87	83.18	91.74
5	MP3	84.23	83.91	92.36
6	Cell phone	85.78	84.61	90.85

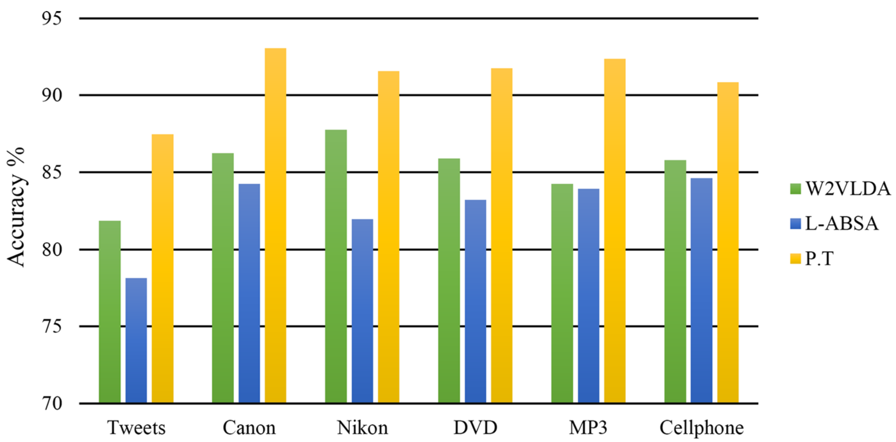

Fig. 6 Comparison results of W2VLDA, LABSA and P.T

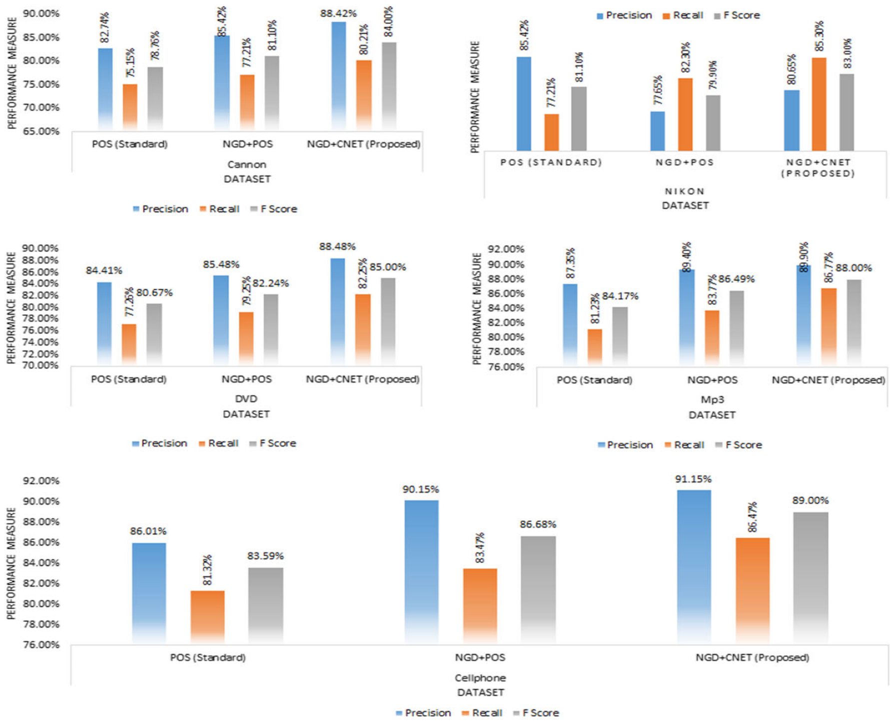

Fig. 7 Results for Canon, Nikon, DVD, MP3 and cell phone datasets

5 Discussion

Our proposed technique shows promising results by using ConceptNet. ConceptNet improves the edge type and edge score transitioning, and hence the results. We also performed experiments on five different product datasets (Canon, Nikon, DVD, MP3, and Cell phone). Figure 7 shows the results of comparisons between traditional Part of Speech Tagging (POS) [47], POS + NGD and proposed approach (NGD + ConceptNet). The subfigures show that our approach performs very well as compared to the existing state-of-the-art methods. The subfigure for Canon dataset shows precision, recall and F-score when above three approaches are applied on Canon. Precision, Recall and F-score of the proposed technique improve 6–7% from POS approach and 3–4% from NGD + POS approach. The subfigure for Nikon dataset shows the results which decreases up to 5–6% from POS approach and 3–4% from NGD + POS approach. The subfigure for DVD dataset shows improvement up to 4–5% from POS approach and 2–3% from NGD + POS approach. The subfigure for MP3 dataset shows the improvement of 2–3% from POS approach and 0.5–1% from NGD + POS approach. Promising results are obtained when the proposed approach is applied to cell phone dataset which gives improvement of 5–6% from POS approach and 1–2% from NGD + POS approach.

6 Conclusion and future work

In this paper, we have presented an aspect identification approach based on aspect co-occurrence and associative similarity. We have applied the proposed technique to each word in an opinionated sentence to consider it either aspect or non-aspect word. We also developed a set of part of speech tagger and integrated it with the proposed technique to obtain significantly better accuracy than existing approaches. The experimental results show that within an appropriate experimental setting, the proposed approach performs better than existing state-of-the-art methods. In future, we can utilize this technique for sentiment analysis problem. Another effort can be made to combine existing techniques with the machine learning techniques for multi-aspect extraction.

References

1. Becker H, Naaman M, Gravano L (2009) Event identification in social media. In: Proceedings of the ACM SIGMOD Workshop on the Web and Databases (WebDB'09)
2. Chatterji S, Varshney N, Rahul RK (2017) AspectFrameNet: a frameNet extension for analysis of sentiments around product aspects. J Supercomput 73(3):961–972
3. Pouriyeh S, Allahyari M, Kochut K, Cheng G, Arabnia HR (2018) Combining word embedding and knowledge-based topic modelling for entity summarization. In: 2018 IEEE 12th International Conference on Semantic Computing (ICSC), pp 252–255
4. Ali Khan A, Ali Chaudhry I (2015) Object oriented case representation for CBR application in structural analysis. Appl Artif Intell 29(4):335–352
5. Akhgar B, Saathoff GB, Arabnia HR, Hill R, Staniforth A, Bayerl PS (2015) Application of big data for national security: a practitioner's guide to emerging technologies. Butterworth-Heinemann, Oxford
6. Allahyari M, Pouriyeh S, Kochut K, Arabnia HR (2017) A knowledge based topic modeling approach for automatic topic labeling. Int J Adv Comput Sci Appl 8(9):335–349
7. Pouriyeh S, Allahyari M, Kochut K, Cheng G, Arabnia HR (2017) Es-lda: Entity summarization using knowledge-based topic modeling. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), vol 1, pp 316–325
8. Khan AA, Chaudhry IA, Sarosh A (2014) Case based reasoning support for adaptive finite element analysis: mesh selection for an integrated system. Appl Phys Res 6(3):21
9. Kim SM, Hovy E (2007) Crystal: analyzing predictive opinions on the web. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)
10. Abirami AM, Askarunisa A (2017) Sentiment analysis model to emphasize the impact of online reviews in healthcare industry. Online Inf Rev 41:471–486
11. Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 168–177
12. Ding X, Liu B, Yu PS (2008) A holistic lexicon-based approach to opinion mining. In: Proceedings of the 2008 International Conference on Web Search and Data Mining, pp 231–240
13. Popescu AM, Etzioni O (2007) Extracting product features and opinions from reviews. In: Kao A, Poteet SR (eds) Natural language processing and text mining. Springer, Berlin, pp 9–28
14. Blair-Goldensohn S, Hannan K, McDonald R, Neylon T, Reis GA, Reynar J (2008) Building a sentiment summarizer for local service reviews. In: WWW Workshop on NLP in the Information Explosion Era, pp 339–348
15. Hofmann T (1999) Probabilistic latent semantic analysis. In: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, pp 289–296
16. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3(1):993–1022
17. Titov I, McDonald R (2008) Modeling online reviews with multi-grain topic models. In: Proceedings of the 17th International Conference on World Wide Web, pp 111–120

18. Zhao WX, Jiang J, Yan H, Li X (2010) Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp 56–65
19. He W, Tian X, Tao R, Zhang W, Yan G, Akula V (2017) Application of social media analytics: a case of analyzing online hotel reviews. Online Inf Rev 41(7):921–935
20. Abirami AM, Askarunisa A (2017) Sentiment analysis model to emphasize the impact of online reviews in healthcare industry. Online Inf Rev 41:471–486
21. Moore MT (2017) Constructing a sentiment analysis model for LibQUAL + comments. Perform Meas Metr 18(1):78–87
22. Agarwal B, Poria S, Mittal N, Gelbukh A, Hussain A (2015) Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach. Cognit Comput 7(4):487–499
23. Poria S, Cambria E, Winterstein G, Huang GB (2014) Sentic patterns: dependency-based rules for concept-level sentiment analysis. Knowl Based Syst 69:45–63
24. Liu Q, Gao Z, Liu B, Zhang Y (2015) Automated rule selection for aspect extraction in opinion mining. IJCAI 15:1291–1297
25. Zheng X, Lin Z, Wang X, Lin KJ, Song M (2014) Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification. Knowl Based Syst 61:29–47
26. Ravi Kumar V, Raghuveer K (2013) Dependency driven semantic approach to product features extraction and summarization using customer reviews. In: Meghanathan N, Nagamalai D, Chaki N (eds) Advances in computing and information technology. Springer, Berlin, Heidelberg, pp 225–238
27. Qiu G, Liu B, Bu J, Chen C (2011) Opinion word expansion and target extraction through double propagation. Comput Linguist 37(1):9–27
28. Samha AK (2016) Aspect-based opinion mining using dependency relations. Int J Comput Sci Trends Technol (IJCST) 4:113–123
29. Sahraei S, Alizadeh MR, Talebbeydokhti N, Dehghani M (2018) Bed material load estimation in channels using machine learning and meta-heuristic methods. J Hydroinform 20(1):100–116
30. Lu Y, Zhai C, Sundaresan N (2009) Rated aspect summarization of short comments. In: Proceedings of the 18th International Conference on World Wide Web pp 131–140
31. Zhao WX, Jiang J, Yan H, Li X (2010) Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing pp 56–65
32. https://dev.twitter.com/overview/api. Accessed 01 Jan 2018
33. http://metashare.ilsp.gr:8080/. Accessed 01 Jan 2018
34. Andreeva OGV, Bagirov MB, Dankina AA, Fedorova TYO, Sheveleva MM (2018) Intellectual analysis of data on the basis of Stanford CoreNLP for pos tagging of texts in the Russian language. Sistemy i Sredstva Informatiki, Syst Means Inform 28(2):145–153
35. Shelke N, Deshpande S, Thakare V (2017) Domain independent approach for aspect oriented sentiment analysis for product reviews. In: Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications pp 651–659
36. Cilibrasi RL, Vitanyi PM (2007) The google similarity distance. IEEE Trans Knowl Data Eng 19(3):370
37. Ritze D, Bizer C (2017) Matching web tables to DBpedia-a feature utility study. Context 42(41):19
38. Bastide F, Sérandat I, Gombert J, Laurent E, Morel E, Kolopp J, Poupard P (2017) Characterization of fungal pathogens (Diaporthe angelicae and D. eres) responsible for umbel browning and stem necrosis on carrot in France. Plant Pathol 66(2):239–253
39. Kunkel J, Loepp B, Ziegler J (2018) Understanding latent factors using a GWAP. arXiv preprint arXiv:1808.10260
40. Wu C, Wu F, Wu S, Yuan Z, Huang Y (2018) A hybrid unsupervised method for aspect term and opinion target extraction. Knowl Based Syst 148:66–73
41. García-Pablos A, Cuadros M, Rigau G (2018) W2VLDA: almost unsupervised system for aspect based sentiment analysis. Expert Syst Appl 91:127–137
42. De Kok S, Punt L, van den Puttelaar R, Ranta K, Schouten K, Frasincar F (2018) Level Aspect-based sentiment analysis using ontology. In: SAC, Pau France. https://doi.org/10.1145/3167132.3167163

Affiliations

Asif Nawaz[1] · **Sohail Asghar**[2] · **Syed Husnain Abbas Naqvi**[1]

Sohail Asghar
sohail.asg@gmail.com

Syed Husnain Abbas Naqvi
husnain.naqvi@iiu.edu.pk

1 Department of Computer Science and Software Engineering, International Islamic University, Islamabad, Pakistan

2 Department of Computer Science, COMSAT Institute of Information Technology, Islamabad, Pakistan