Asian Journal of Information Technology 17 (3): 212-217, 2018
ISSN: 1682-3915

Opinion Mining with Aspects and Shortcuts for Prediction Model

[1]Ghada Hafez, [1]Rasha Ismail and [2]Omar Karam
[1]Department of Information Systems, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
[2]Department of Information Systems, Faculty of Computer and Information Sciences, British University of Egypt, Cairo, Egypt

Abstract: Opinion mining or sentiment analysis extract specified information from a large amount of text or reviews given by the internet users. Opinion mining classifies the large text of opinions as positive (good), negative (bad) or neutral. According to the number of positive, negative and neutral reviews, the product or service will be rated. Most researches neglected the shortcuts of words like (CD "Compact Disc"), (GR8 "means Great"). To avoid classify these words wrong, a database is built for some of these shortcuts with its meaning and orientation (positive, negative or neutral). Also, few researches tried to extract tempo words from text as general not from opinions, so that, a framework is proposed to merge tempo and sentiment analysis to enhance the prediction for the opinions.

Key words: Aspect mining, opinion mining, tempo wrods, text mining, shortcuts, database

INTRODUCTION

Opinion mining is a field of web content mining that aims to find valuable information out from users opinions. Mining opinions on the web is a fairly new subject and its importance has grown significantly mainly due to the fast growth of e-Commerce, blogs and forums (Dhokrat *et al*., 2015, 2016). On the other hand, temporal opinion mining attracted a big attention in the fields of Natural Language (NL) and social networks. TempoWordNet is one of the attempts to building a lexicon that helps in finding temporal synsets. It contains words extracted from wordnet, in addition to probabilities of each word whether it is "future", "present", "past" or "atemporal". Temporal classifiers are learn from a set of time sensitive synsets and then are incorporated to Wordnet to form TempoWordNet. So, each synset is augmented with its intrinsic temporal value (Dias *et al*., 2014).

Mobile application opinions by customers contain an abundance of data on the issues that customers are encountering. Developers, clients and the application store owners (the product organization) can profit by a superior comprehension of these issues-developers can better comprehend customer's worries, application store proprietors can center around enhancing the item and customers can contrast comparable applications with choose which ones to download or buy. Be that as it may, customer reviews are not named, e.g., we don't know which kinds of issues are brought up in a survey. Subsequently, one must filter through possibly a great many surveys with slang and truncations to comprehend the different kinds of issues. Additionally, the unstructured and casual nature of reviews complicates the automated labelling of such reviews (McIlroy *et al*., 2016). Moreover, Twitter is an online social communication service where overall customers distribute their assessments on an assortment of points, talk about current issues, complain and express positive or negative supposition (sentiment) for items they use in day by day life. Consequently, Twitter is a rich source of information for assessment mining and opinion examination. In any case, estimation examination for Twitter messages (Tweets) is viewed as a problem issue since tweets are short and casual (Wolny, 2016). As the number of Internet and mobile phone users increase, texting and chatting have become popular ways of communication. The extensive use of mobiles and Internet led into the creation of a new language where words are transformed and made shorter using different styles. Shortcut texting is used in informal areas such as SMS, online reviews, chat rooms, forums and posts in social networks. Large amounts of data founded from these informal sources can be utilized for various tasks in machine learning and data analytics. These data may be written in shortcut forms, text normalization is necessary before NLP actions such as information extraction, data mining, text summarization, opinion classification can be fully achieved by acting as a preprocessing that transforms all informal texts back to their original and understandable forms (Nocon *et al*., 2014).

Corresponding Author: Ghada Hafez, Department of Information Systems, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt

Using sentiment analysis which becomes one of the most important sources in decision making can help us to extract, identify, evaluate or otherwise characterizes from the online sentiments reviews. Although, bag-of-words Model is the most widely used technique for sentiment analysis, it has two major weaknesses, using a manual evaluation for a lexicon in determining the evaluation of words and analyzing sentiments with low accuracy because of neglecting the language grammar effects of the words and ignore semantics of the words. El-Din (2016) In this study, a new system is proposed.

MATERIALS AND METHODS

By McIlroy *et al*. (2016), they studied the multi labelled nature of reviews from 20 mobile apps. in the Google Play Store and Apple App. Store. They found that up to 30% of the reviews raise various types of issues in a single review (e.g., a review might contain a countenance request and a bug report). They proposed an approach that can automatically assign multiple labels to reviews based on the raised issues with a precision of 66% and recall of 65%. Then they applied their approach to address three proof-of-concept analytics use case scenarios, comparing competing apps to assist developers and users, providing an overview of 601,221 reviews from 12,000 Apps. in the Google Play Store to assist app. store owners and developers and detecting anomalous Apps. in the Google Play Store to assist app store owners and users. By Wolny (2016) they focused on the issue of the breaking down of images called feeling tokens, including feeling images (e.g., emojis and emoticon ideograms). As indicated by perception, these feeling tokens are usually utilized. They specifically express one's feelings despite his/her dialect, subsequently they have turned into a helpful flag for estimation examination on multilingual tweets. The study depicts the way to deal with performing sentiment analysis that can decide positive, negative and impartial (neutral) sentiment for a tested point. By Nocon *et al.* (2014) they introduced the NormAPI (text analysis, application program interfaces, learning (artificial intelligence), natural language processing, pattern classification" INSPEC, controlled indexing" and "INSPEC, non-controlled Indexing "bilingual translation, NormAPI, application program interface, filipino shortcut texts, texting communication, chatting communication, social networks, machine learning, data analytics, text normalization, NLP action, natural language processing, information extraction, data mining, text summarization, opinion classification, an API for normalizing Filipino shortcut texts. NormAPI primarily intends to be used as a preprocessing system that corrects informalities in shortcut texts before they are handed for complete data processing. Users created implants as the modification of languages such as shortcuts (e.g. , 4 get) where texts are shortened for faster replies or posts. These informal or shortcut texts are not unique in the Filipino language. In addition, shortcuts are derived from several styles and they vary per country. By El-Din (2016) The proposed method depends on the improvement bag of-words demonstrate for assessing opinion extremity and score consequently by utilizing the words weight rather than term recurrence. This system like wise can group the reviews in light of features and key words of the logical theme area. This study presents answers for fundamental sentiment analysis challenges that are appropriate for the review structure. It like wise looks at the impacts by the proposed improvement model to achieve higher exactness. By Dias *et al.* (2014) they introduced tempowordnet which is built based on WordNet such in which each synset is automatically time-tagged with four dimensions, temporal, past, present and future each synset is associated to its intrinsic temporal value that extends the basic opinion mining framework to use Reference Time (RT) information in the formation of future predictions. To extract RT, four features are extracted to discriminate prediction opinions with different RT tags.

In the previous reesrach they didn't take in consideration the combination between opinions about aspects of certain product and its time. So, in our research we proposed a system to merge them together using Extracting Verbs algorithm (EV) which will improve the prediction of opinions through time dimensions (past, present and future) depends on tempowordnet. Also, we adding the sentiment analysis for shortcuts and emojis to increases the weight of the opinion.

Overview on the tempowordnet: Tempowordnet is a temporal ontology which may contribute to the success of time-related applications. Temporal classifiers are learned from a set of time sensitive synsets and then applied to the whole WordNet to give rise to TempoWordNet. So, each synset is augmented with its intrinsic temporal value. To evaluate TempoWord-Net, a semantic vector space representation is used for sentence temporal classification which shows that improvements may be achieved with the time-augmented knowledge. There are some problems were deduced when applying it on the opinions.

Table 1: Statistics of the dataset

Variables	Values
Past	2508
Present	820
Future	13758
Atemporal	100512
Total	117598

p>0.5

Most of the words are atemporal, Table 1 shows statistics of the dataset and apparently there are about 79% of the words are detected as atemporal and this is a high percentage. When looking at the words detected as atemporal, we found that they are false positives. So, the atemporal words could be classified to past or present or future. It noticed that the total number of words are 117568, verbs are about 13766 records. Number of atemporal verbs are 12058 which representing 10.2% from total atemporal words.

Extracting Verbs (EV) algorithm: In our previous research (temporal sentiment analysis and time tags for opinions) we proposed a new algorithm extracting verb. This algorithm is for solving problems mentioned above (an overview on tempowordnet), it was noticed that most of these words are verbs in their simple tense, so, we worked on this point to enhance the detection of the tense of the verb. According to the mining process, each word was checked and determine whether it is a verb or not, the following steps were followed:

Extracting Verbs (EV) algorithm:
If the verb ended with (ed) It classified as (past). But if the sentence contains any of those words "has /have/just/already/yet/ever/never" so, the verb will classified as (present)
2) if the verb ended with (ing)
It classified as (present)
if the sentence contains any of those words "am/is/are" and "look/listen /now/at the moment/at the present/today-still—hurry up-please–don't"
And if the sentence contains "was/were" So, the verb will classified as (past))
And it will be (future) if the sentence contains soon–tomorrow–in the future-tonight–next-in a few minutes–in the evening-shortly-here after-In a little time-In the years to come-In future-early-later-today evening-with in a week
3) if the verb is in the source/simple format
It classified as (future) if the sentence contains any of the modals verbs "can, could, may, might 3, will, would, shall, should, must, ought to"
And (present) if none of those words are exists
4) Irregular verbs
If the verb come as the first column, so, it follows the condition number 3
If as column 2 or 3 so it is (past)

The proposed system architecture: We proposed a temporal opinion mining system which provides users a quality aspect-based summary and the orientation of the opinion" past, present, future, atemporal", out of product

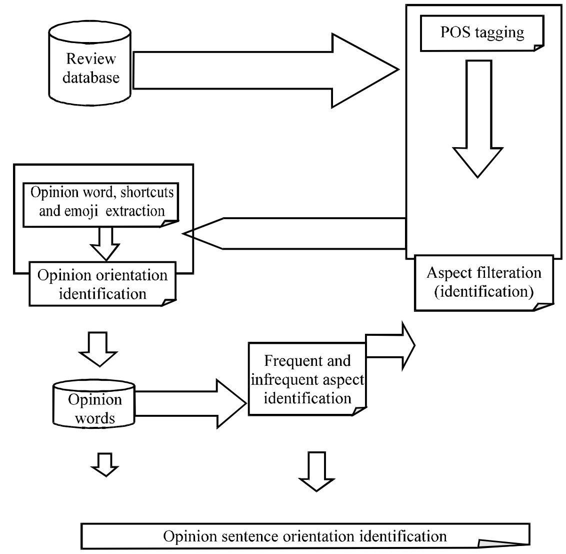

Fig. 1: Proposed system architecture

opinions from ordinary customers on the web. The architecture consists of 4 modules as in Fig. 1 (Dias *et al.*, 2014).

Part-of-speech tagging: A Part-of-Speech tagger (POS tagger) was used to characterize limits (split conclusions into sentences) and to create for each word a given grammatical feature. In aspect identification, data mining framework will rely upon the noun or noun phrases (two up to three neighbor nouns in a sentence) created on this progression to deliver various incessant aspects.

Opinion word, shortcuts and opinion sentence: An opinion word is a term used to indicate a word that usually populate an object or an attribute of this object. They are usually adjectives and adverbs but they can also be nouns and verbs. An opinion sentence is a sentence which holds at least one reference to the object (that could be the object itself or any attribute of the object) and also includes one or more opinion words.

The sentence "GR8, I like it very much", word "GR8" means "great" in the lexion and so will be classified as a positive word and "like" also. So, this sentence will be classified as positive at all.

Aspect identification: Aspect identification is the process used to deduce possible product aspects out of the tagged texts generated by the last step. The part of-speech responsible for giving names to entities of the real world are nouns in this case a noun gives name to the product and its aspects (i.e., screen, battery, camera, etc.). In these reserach, they define two categories of aspects, frequent aspects and infrequent aspects.

The biggest downside is that the output (the frequent aspects) depend a lot on the number of opinions being analyzed. Also, there is no guarantee that a frequent aspect found by the system is actually a real aspect.

Frequent and infrequent aspects identifcation: Aspects that appear on many opinions have more chance to be relevant and therefore, more likely to be actually a real product aspect. Examples:

- The camera is wonderful
- The applications of the mobile is wonderful

In the above example, the two sentences have one opinion word in common, "wonderful". Because an opinion word could be used to describe more than one object, these opinion words are used to look for aspects which couldn't be found in the step described.

Proposed system achieves three goals, prepare an aspect-based summary from the opinions retrieved in 1 which consists of product name, the number of opinions analyzed and each identified aspect (ex, camera, screen, voice quality, ..., etc.,). Then, for each found identified aspect, the system must classify the sentiment associated with each of them (either negative or positive), along with text fragments where these classified aspects were mentioned. Extract shortcuts words and emojis and classify them according to the ready building database as (positive, negative or neutral) and the orientation of the opinion (past, present, future and atemporal).

For example, if the orientation of a certain opinion is negative and past, this may indicates that it is now better. And if the orientation is negative and present, this is an important indicator that this aspect of the product need to be improved and so on.

Datsa set: For evaluation, we use mobile reviews with the search terms "mobile", "phone" and "cellular". The reviews are retrieved between April 14th and 18th, 2017 and include only the mobile. The resulting list contains 1000 reviews (McIlroy *et al*., 2016).

Evaluation measures: We compare the results of the classifier under test with trusted external judgments. The terms positive and negative refer to the classifier's prediction and the terms true and false refer to whether that prediction corresponds to the external judgment.

Recall in this context is also referred to as the true positive rate or sensitivity and precision is also referred to as Positive Predictive Value (PPV), other related measures used in classification include true negative rate and accuracy. True negative rate is also called specificity.

RESULTS AND DISCUSSION

Table 1 shows the entities of the results, Table 2 as shown represents the results without the shortcuts. Table 3 shows the results after adding shortcuts. Later in Table 4 the comparison between the two results were summarized through the cross validation process using the entire annotated dataset. The accuracy after the adding the shortcuts was 0.794705561, 0.940113 and 0.879842 for total past, present and future, respectively which is better than before adding shortcuts which was was 0.72337705, 0.7862134 and 0.694286.

The F-measure, a metric which correlates precision and recall was also better after adding the shortcuts. Once, the sentiments classifier has been built, validated and also used all the datasets, the next step was to apply the general semantic orientation of the sentiments. Table 5 shown the results for both two results (Fig. 2 and 3).

Table 5 and 6 shows that after adding shortcuts the results are much better that is because the number of words adding in the classification of the opinions are increased (Fig. 4 and 5).

Table 2: Example of the irregular verbs

V1 base form	V2 past simple	V3 past participle
Awake	Awoke	Awoken
Be	Was, were	Been

Table 3: Example of the shortcuts and its meaning

Shortcut	Meanings
LOL	Laughing out Loud
GR8	Great
BBL	Be Back Later
ASAP	As Soon as Possible
B4	Before
B4N	Bye for Now
Def	Definitely
tmrw	Tomorrow
ETA	Estimated Time of Arrival
IDC	I Don't Care
THX	Thanks
PLS	Please

Here, we have a shortcut like "tmrw", we will take the meaning of it and will be classify as future

Table 4: Example of the emojis and its meaning and the sentiment of each

Emojis	Meanings	Sentiments
:-)	Basic smiley	Positive
:)	Midget smiley	Positive
,-)	Winking happy smiley	Positive
(:-(	Very unhappy smiley	Negative
8-O	Oh my god	Positive/negative
:-C	Real unhappy smiley	Negative
:-e	Disappointed smile	Negative

Table 5: Sample of the results without shortcuts

ID	Aspect	Past positive	Past negative	Present positive	Present negative	Future positive	Future negative	Count
1	Screen	0	0	13	0	7	3	23
2	GPS	0	0	0	3	0	0	3
3	Storage	0	0	22	0	0	0	22
4	Network	0	0	0	5	0	0	5
5	battery	0	11	0	0	8	0	19

Table 6: Sample of results with shortcuts

Aspect	Past positive	Past negative	Present positive	Present negative	Future positive	Future negative	Count
Screen	0	0	13	0	7	3	23
GPS	0	0	0	3	0	0	3
Storage	0	0	22	0	0	0	22
Network	0	0	0	5	0	0	5
battery	0	11	0	0	8	0	19
shortcuts	3	0	1	2	0	0	6

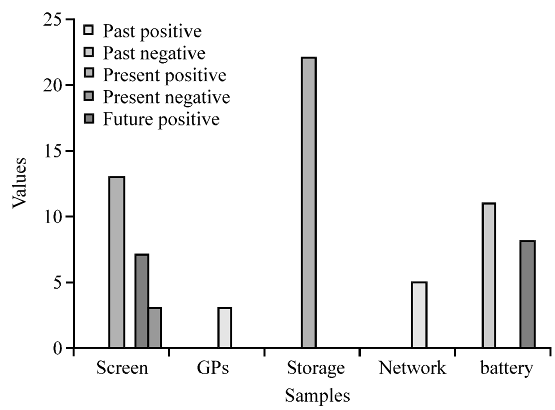

Fig. 2: Representing a sample of the results without shortcuts

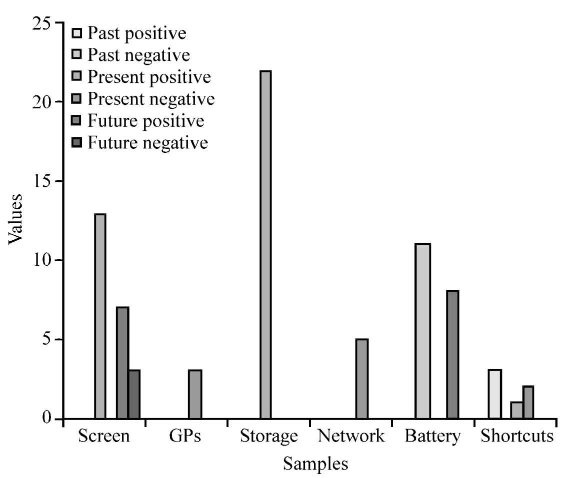

Fig. 3: A sample of the results with shortcuts

In this study, we propose a framework for merging temporal orientation references (future, past and present) opinions with sentiment analysis for aspects and shortcuts of words. The goal of the system is merging the opinion mining with time. Not just classify them as (positive, negative) but also as (past, present, future and atemporal), this will be useful in prediction. For example, if the orientation of a certain opinion is negative and past, this may indicates that it is now better and if the orientation is negative and present, this is an important indicator that this aspect of the product need to be improved (Table 7 and 8).

Adding shortcuts of words to our system increases the weight of the orientation of the opinion as positive or negative.

Table 7: Comparison between the two results

	Percentage					
Accuracy	Total past (positive)	Total past (negative)	Total present (positive)	Total present (negative)	Total future (positive)	Total future (negative)
After shortcuts	79.47	78.35	84.01	80.52	82.98	69.22
Before shortcuts	75.33	73.42	82.62	79.62	79.42	68.34

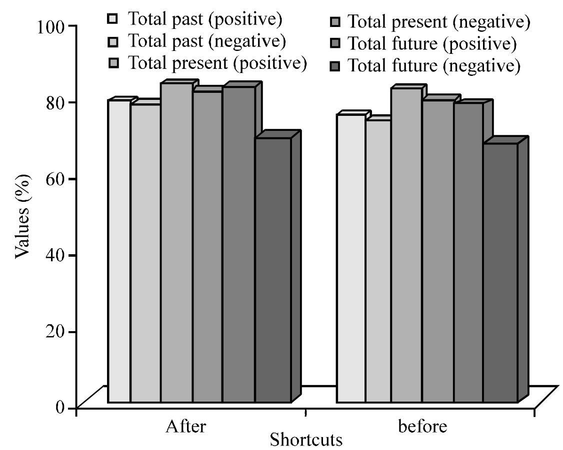

Fig. 4: A comparison between the two results

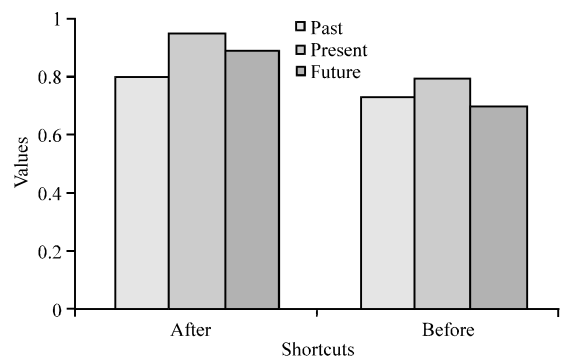

Fig. 5: A comparison between the two results

Table 8: Shows the comparison between the two results

	Before shortcuts			After shortcuts		
The	Total past	Total present	Total future	Total past	Total present	Total future
Precision	0.80555556	0.7002351	0.930894	0.8210021	0.8909091	0.96
Recall	0.83205385	0.97632	0.825825	0.78530769	0.852387	0.812077
F1-measure	0.78251959	0.84438889	0.765881	0.80007182	0.8556276	0.831165
Accuracy	0.72337705	0.7862134	0.694286	0.794705561	0.940113	0.879842

CONCLUSION

The goal of the system is merging the opinion mining with time. Not just classify them as (positive, negative) but also as (past, present, future and atemporal), this will be useful in prediction. For example if the orientation of a certain opinion is negative and past, this may indicates that it is now better and if the orientation is negative and present, this is an important indicator that this aspect of the product need to be improved and so on. Also, extracting the shortcuts words and emojis according to a ready made lexicon to increase the weight of the positive or negative opinions.

RECOMMENDATIONS

For future research, we will build a more complex system which analyzes the opinion sentence with the more complex grammar.

ACKNOWLEDGEMENT

Many thanks to the supervisors who helps me to make this research.

REFERENCES

Dhokrat, A., S. Khillare and C.N. Mahender, 2015. Review on techniques and tools used for opinion mining. Intl. J. Comput. Appl. Technol. Res., 4: 419-424.

Dhokrat, A., S. Khillare and C.N. Mahender, 2016. Review on opinion mining for fully fledged system. Indonesian J. Electr. Eng. Inf., 4: 141-148.

Dias, G.H., M. Hasanuzzaman, S. Ferrari and Y. Mathet, 2014. Tempowordnet for sentence time tagging. Proceedings of the 23rd International Conference on World Wide Web, April 7-11, 2014, ACM, Seoul, Korea, ISBN: 978-1-4503-2745-9, pp: 833-838.

El-Din, D.M., 2016. Enhancement bag-of-words model for solving the challenges of sentiment analysis. Intl. J. Adv. Comput. Sci. Appl., 7: 244-247.

McIlroy, S., N. Ali, H. Khalid and A.E. Hassan, 2016. Analyzing and automatically labelling the types of user issues that are raised in mobile app reviews. Empirical Software Eng., 21: 1067-1106.

Nocon, N., G. Cuevas, D. Magat, P. Suministrado and C. Cheng, 2014. NormAPI: An API for normalizing Filipino shortcut texts. Proceedings of the 2014 International Conference on Asian Language Processing (IALP), October 20-22, 2014, IEEE, Kuching, Malaysia, ISBN:978-1-4799-5331-8, pp: 207-210.

Wolny, W., 2016. Sentiment analysis of Twitter data using emoticons and emoji ideograms. Studia Ekonomiczne, 296: 163-171.