Scientific
Research
Publishing

Hybrid Algorithm to Evaluate E-Business Website Comments

Osama M. Rababah[1], Ahmad K. Hwaitat[2], Dana A. Al Qudah[1], Rula Halaseh[1]

[1]Department of Business Information Technology, The University of Jordan, Amman, Jordan
[2]Department of Computer Science, The University of Jordan, Amman, Jordan
Email: o.Rababah@ju.edu.jo

Abstract

Online reviews are considered of an important indicator for users to decide on the activity they wish to do, whether it is watching a movie, going to a restaurant, or buying a product. It also serves businesses as it keeps tracking user feedback. The sheer volume of online reviews makes it difficult for a human to process and extract all significant information to make purchasing choices. As a result, there has been a trend toward systems that can automatically summarize opinions from a set of reviews. In this paper, we present a hybrid algorithm that combines an auto-summarization algorithm with a sentiment analysis (SA) algorithm, to offer a personalized user experiences and to solve the semantic-pragmatic gap. The algorithm consists of six steps that start with the original text document and generate a summary of that text by choosing the N most relevant sentences in the text. The tagged texts are then processed and then passed to a Naïve Bayesian classifier along with their tags as training data. The raw data used in this paper belong to the tagged corpus positive and negative processed movie reviews introduced in [1]. The measures that are used to gauge the performance of the SA and classification algorithm for all test cases consist of accuracy, recall, and precision. We describe in details both the aspect of extraction and sentiment detection modules of our system.

1. Introduction

The apparent revolution of the user-generated content in Web 2.0 has enabled people to share and present their knowledge and experience. Web users have enthusiastically combined peer-authored posts, recommendations

and online reviews into their lives. As web users' trust grew in such reviews and recommendations to the point of building their choices upon them, the selection could vary from planning a night out or choosing a movie to watch by checking the reviews of others who previously saw the film. The evolution of Web 2.0 allows everyone to have a voice which assists in increasing the human collaboration capabilities on a worldwide scale, enabling individuals to share opinions by reading and writing web generated contents and user generated contents. Even with the growing acceptance of the new social networking technologies, there has been little research on the quality of the content provided and how it affects other organizations and major marketing decisions. Since websites that offer user reviews can be surprisingly technologically inferior, which will lead to having no choice but to browse through massive amounts of text to find a particular piece of interesting information [2].

Sentiment analysis is the process of determining the polarity or intention of a written text and according to [2]. SA includes five steps to analyze sentiment data. The first step begins with data collection which consists of collecting data from a user-generated content contained in blogs, forums and social media networks. The collected data can be messy and expressed by different methods or by using different words, slangs, and context of writing and manual analysis to such huge amount of data is not possible and exhausting. Therefore, text analytics and natural language processing are used to mine and classify the data. Secondly, it comes to the text preparation step, and it consists of cleaning the extracted data before analysis. Non-textual contents and contents that are inappropriate for the study are recognized and removed at this step. The third step in the process of SA is emotion detection, in which the extracted sentences of the reviews and opinions are inspected; sentences with individual expressions are retained, and sentences with objective communication are discarded. The fourth step is sentiment classification where personal sentences are classified in positive, negative, good, bad, like, dislike, but classification can be made by using multiple points. Finally, it is the presentation of output step where the key objective of sentiment analysis here is to transform unstructured text into meaningful information. At the end of the study, the test results are displayed on graphs like a pie chart, bar chart and line graphs. Also, time can be analyzed and can be graphically displayed a sentiment timeline with the chosen values of frequency, percentages and averages over time [2].

Efficient auto-summarization of texts is a standalone field of study in the computational linguistic community. One of its top goals is to offer users a way to access the content of interest to them in a faster and a more efficient way. SA, on the other hand, aims to be able to divide correctly text data into categories based on the opinions the authors expressed about particular issues, using natural language. To be able to offer personalized user experiences, these two fields can be analyzed in a holistic way. The present paper does that by merging an auto-summarization algorithm with a sentiment analysis algorithm and analyzing the results using the relevant metrics of accuracy, recall, and precision.

An opinion is simply a positive or negative attitude, view, emotion, or appraisal about an entity or an aspect of the entity from an opinion owner at a particular time [3]. Accessing and searching reviews is frustrating when users have a vague idea of the product or its features and they need a recommendation or a close match. Keyword based search does not usually provide good results, as the same keywords can appear in both good and bad reviews [2]. Another challenge in understanding studies is that a reviewer's overall rating might be focused on the product features in which might not be of interest to the user making the search. More challenges include having the sentiment word with an opposite meaning in a particular domain. The use of conditional and interrogative sentences challenge is where SA words are more of a neutral view while in an opinion they can be either positive or negative. Sarcastic sentences may violate the meaning of sentences, therefore, close attention to the words used in such sentences is needed. Other issues include when people write a word in different ways which may not give us an indication that it is the same word (*i.e.* Motorola and Motto). Moreover, sentences can be lacking sentiment words at times but carrying either a positive or negative feedback about a particular topic. Author reader understanding can pose a limitation as well since the reader can misunderstand the author intention at times and another thing which can result in wrong results and statistics which is the spam posts that are added intentionally to give more positive feedback or to destroy the reputation of a certain organization [4]. People methods of expression can be contradictory while most of the traditional text processing methods depend on the fact that a small difference between two pieces of text doesn't alter the meaning.

This paper is organized as follows. Section 2 is a literature review of relevant work. Section 3 gives an overview of the methodology we adopt for both parts of the algorithm. Sections 4 and 5 present test cases and results obtained by running the algorithm. Finally, Section 6 demonstrates the conclusion.

2. Literature Review

The term SA first appeared in [5]. However, the research on sentiments appeared earlier [6]-[10]. The literature on SA concentrated on different fields, from computer science to management sciences, social sciences and business due to its importance to various tasks such as subjective expressions [11], sentiments of words [12], subjective sentences [13], and topics [5] [13] and [14].

SA can be approached in different manners, either by classifying data into two categories: positive or negative [15] or by using various intermediary classes, such as the multiple stars reviews [1]. The sentiment classification approaches can be classified into machine learning, lexicon based and hybrid approach [16] which will be our focus in this paper.

The machine learning approach is used for predicting the polarity of sentiments based on trained as well as test data sets. The lexicon-based approach does not need any prior training to mine the data. Finally in the hybrid approach, the combination of both the machine learning and the lexicon based methods has the potential to increase the sentiment classification performance [2].

The growing in new types of online information also changes the type of summarization that is of interest. Summarization has newly been combined with work on SA [17]-[19]. In this context, it is desirable to have summaries which list the pros and cons of a service or product. For example, for a restaurant, one might want to hear that the service is good, but the food is bad. Given the numerous different reviews that one can find on the web, the problem is to identify common opinions. Some of the approaches that have been tried so far include: determining semantic properties of an object, defining the intensity of an opinion, and determining whether an opinion is important. In this paper, we present a hybrid algorithm that was uniting an auto-summarization algorithm with an SA algorithm to offer personalized user experiences and to solve the semantic- pragmatic gap. A novel aspect of these models is that they exploit user-provided labels and domain specific characteristics to increase personalized user experience.

The auto-summarization of texts was done using the tools offered by the NLTK toolkit (NLTK.org) [20], which provide the opportunity to tag sentences syntactically and calculate word frequencies and perform stop word elimination, by using the pre-defined English corpora.

3. Methodology

The methodology, we employ for both parts of the algorithm–summarization and sentiment analysis will be described in the following subsections.

3.1. Hybrid Auto Summarization Algorithm and Sentiment Analysis Algorithm

The figure below shows the Hybrid Auto-summarization and sentiment analysis algorithm discussed earlier in the introduction.

The auto summarization algorithm consists of 6 steps which start from the original text document that is given as an argument (step 1) and generate a summary of that text by choosing the n most relevant sentences in it (where n is a user-defined variable) (step 6). The intermediate steps (2 - 5) consist of sentence tagging and word frequency and relevance calculations.

The tagged texts are being processed and then passed to a naïve Bayesian classifier, along with their tags as training data. After the classifier has been trained, new comments are given to it for classification. The steps of the auto-summarization and Sentiment analysis algorithm that we use for this paper are found in **Figure 1**.

3.2. Relevant Measures

The measures that we use to gauge the performance of the sentiment analysis and classification algorithm for all test cases consist of the recall, the precision and the accuracy equations illustrated in the below equations but first we define some terms used in the equations [21]:

True positives: positive comments correctly identified as positive.
True negatives: negative comments correctly identified as negative.
False positives: negative comments incorrectly identified as positive.
False negatives: positive comments incorrectly identified as negative.
The sensitivity equation, the true positive rate and sometimes called recall, measures the proportion of posi-
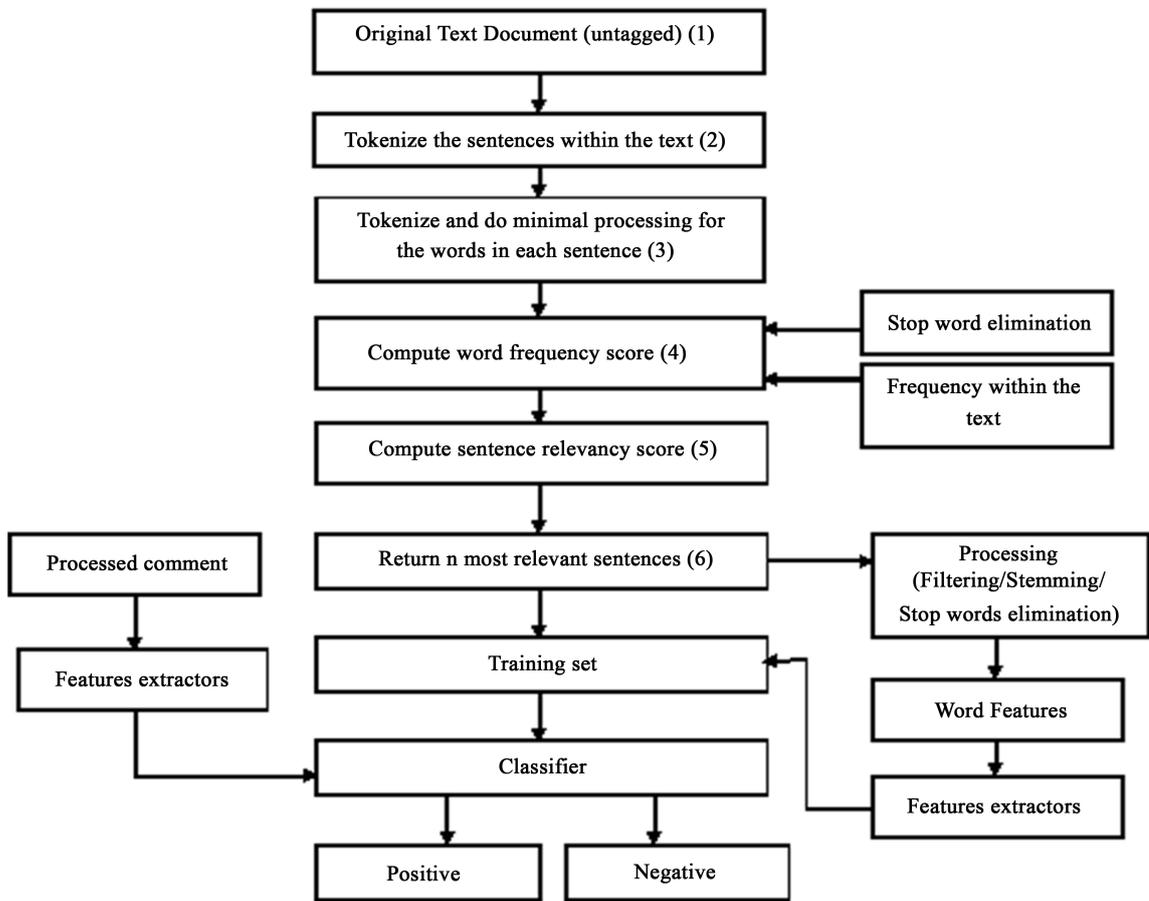
Figure 1. Hybrid auto-summarization and sentiment analysis flowchart.

tives that are correctly identified such as the percentage of positive restaurant or movie reviews that are truly positive. Recall is calculated as below Equation (1):

$$\frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{1}$$

The precision equation or the positive predictive value equation is the fraction of relevant retrieved instances such as the percentage of negative restaurant or movie reviews that are truly negative and we calculate it in Equation (2):

$$\frac{\text{True Negatives}}{\text{True Positives} + \text{False Positives}} \tag{2}$$

Accuracy is defined as the closeness of agreement between the result of a measurement and a true value [22]. With the accuracy equation, we get the rate at which items are correctly classified and/or retrieved. We calculate the accuracy rate in Equation (3):

$$\frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{False Negatives} + \text{True Negatives} + \text{False Positives}} \tag{3}$$

The values that were obtained for each of these indicators are shown and discussed in the results section of this paper.

4. Test Cases

For this article we have five different types of test cases:

- *No Proc process* which uses the original texts of the comments for both training and classification, with the dataset divided 20%/80%.
- *Min Proc process* only eliminates punctuation and uppercase letters, still uses the original complete textual comments for classification.
- *Sum on Sum* where all comments are summarized first and then they are used for training the Bayesian network and testing (again the 80%/20% ratio was used for the classification/testing).
- *Sum on full* where the Bayesian network is trained with the full text of the comments and the summaries are given as new items to be classified.
- *Full on Sum* where the Bayesian network is trained with the text of the summaries and the full textual comments texts are used for classification.

The 20%/80% ratio for training vs. classification was respected for all test cases, and no text was used for both training and testing (a summary of a text is considered the equivalent of the original text in this regard).

These test cases were devised so that a clear picture of how auto summarization affects the accuracy of a sentiment analysis algorithm, having the original classification (No Proc) as a baseline.

5. Results

The results we obtained from running the algorithm for each of the test cases are shown below.

We can observe from both **Table 1** and **Figure 2** that the best metrics are obtained for the Sum on Sum, Min-Proc, and NoProc. This result will be discussed further in the conclusion section of the paper.

We can also be observed in **Figure 3** that the best accuracy is obtained for the NoProc and MinProc.
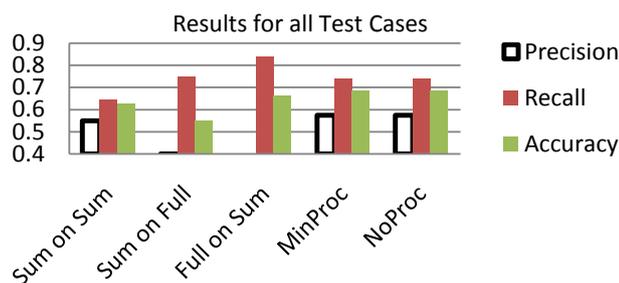
Figure 2. Complete value set for all test cases.
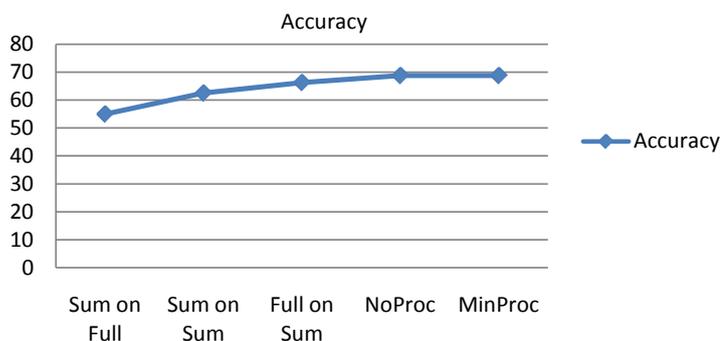
Figure 3. Accuracy depending on the test case.

Table 1. Numeric values for all the test cases.

	Precision	Recall	Accuracy
Sum on Full	0.15	0.75	0.55
Sum on Sum	0.55	0.65	0.63
Full on Sum	0.4	0.84	0.66
NoProc	0.58	0.74	0.69
MinProc	0.58	0.74	0.69

Also, we notice that, in comparison with the other metrics, accuracy has the smallest variation amongst the test cases.

6. Discussions and Conclusions

After running the algorithm against the tagged data, we conclude that the accuracy of the algorithm is better when texts of the same type are used for training and testing (summaries tested with summaries and long texts with long texts), showing that the summaries, by using fewer words affect the pragmatic of the texts. Despite the previous point, the accuracy of the algorithm does not vary greatly between test cases—as opposed to the other metrics. There was no difference in the results that were obtained for texts that were not processed at all and the ones that had undergone minimal processing, showing that polarity was not influenced by the usage of upper case vs. lower case or punctuation signs (at least in the corpus under study). The precision drops drastically when the algorithm is trained with different types of texts than the ones they are trained with (Sum on Full and Full on Sum test cases), the explanation being the same as the one for the accuracy drop.

As further developments for the proposed algorithm, the following directions could be investigated: 1) Variation of the number of sentences in the summaries depending on the length of the original text—assuring that the length of the original text does not affect the training algorithm; 2) Extra processing methods could be added to the algorithm, such as stemming and stop word elimination, and the results should be reexamined to determine if the performance metrics improve for the mix test cases—Sum on Full and Full on Sum.

References

[1] Pang, B. and Lee, L. (2004) A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. *Proceedings of the* 42*nd Annual Meeting on Association for Computational Linguistics*. http://dx.doi.org/10.3115/1218955.1218990

[2] D'Andrea, A., Ferri, F., Grifoni, P. and Guzzo. T. (2015) Approaches, Tools and Applications for Sentiment Analysis Implementation. *International Journal of Computer Applications*, **125**, 26-33.

[3] Abdulla, N., Ahmed, N., Shehab, M., AlAyyoub, M., Al-Kabi, M. and Al-Rifai, S. (2014) Towards Improving the Lexicon-Based Approach for Arabic Sentiment Analysis. *International Journal of Information Technology and Web Engineering* (*IJITWE*), **9**, 55-71. http://dx.doi.org/10.4018/ijitwe.2014070104

[4] Sindhu, R., Jamail, R. and Kumar, R. (2014) A Novel Approach for Sentiment Analysis and Opinion Mining. *International Journal of Emerging Technology and Advanced Engineering*, **4**, 522-527.

[5] Nasukawa, T. and Yi, J. (2003) Sentiment Analysis: Capturing Favorability Using Natural Language Processing. *Proceedings of the* 2*nd International Conference on Knowledge Capture*, Florida, 23-25 October 2003, 70-77. http://dx.doi.org/10.1145/945645.945658

[6] Morinaga, S., Yamanishi, K., Tateishi, K. and Fukushima, T. (2002) Mining Product Reputations on the Web. *Proceedings of the* 8*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 341-349. http://dx.doi.org/10.1145/775047.775098

[7] Pang, B., Lee, L. and Vaithyanathan, S. (2002) Thumbs up? Sentiment Classification Using Machine Learning Techniques. *Proceedings of the* 7*th Conference on Empirical Methods in Natural Language Processing*, 79-86.

[8] Tong, R.M. (2001) An Operational System for Detecting and Tracking Opinions in On-Line Discussion. *Proceedings of SIGIR Workshop on Operational Text Classification*.

[9] Turney, P. (2002) Thumbs up or Thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the* 40*th ACL*, 417-424.

[10] Wiebe, J. (2000) Learning Subjective Adjectives from Corpora. *Proceedings of National Conference on Artificial Intelligence*.

[11] Wilson, T., Wiebe, J. and Hoffmann, P. (2009) Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis. *Computational Linguistics*, **35**, 399-433. http://dx.doi.org/10.1162/coli.08-012-R1-06-90

[12] Hatzivassiloglou, V. and McKeown, K.R. (1997) Predicting the Semantic Orientation of Adjectives. *Proceedings of the* 8*th Conference on European Chapter of the Association for Computational Linguistics Madrid*, Spain, 174-181.

[13] Yi, J., Nasukawa, T., Niblack, W. and Bunescu, R. (2003) Sentiment Analyzer: Extracting Sentiments about a Given Topic Using Natural Language Processing Techniques. *Proceedings of the* 3*rd IEEE International Conference on Data Mining* (*ICDM* 2003), Florida, 19-22 November 2003, 427-434. http://dx.doi.org/10.1109/icdm.2003.1250949

[14] Hiroshi, K., Tetsuya, N. and Hideo, W. (2004) Deeper Sentiment Analysis Using Machine Translation Technology.

Proceedings of the 20*th International Conference on Computational Linguistics* (*COWLING* 2004), Geneva, 23-27 August 2004, 494-500. http://dx.doi.org/10.3115/1220355.1220426

[15] Govindarajan, M. (2013) Sentiment Analysis of Movie Reviews Using Hybrid Method of Naive Bayes and Genetic Algorithm. *International Journal of Advanced Computer Research*, **3**, 139-145.

[16] Albanese, M. (2013) A Multimedia Recommender System. *ACM Transactions on the Internet Technology* (*TOIT*), **13**, 1-32. http://dx.doi.org/10.1145/2532640

[17] Branavan, S., Chen, H., Eisenstein, J. and Barzilay, R. (2008) Learning Document-Level Semantic Properties from Free-Text Annotations. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 263-271.

[18] Carenini, G. and Cheung, J. (2008) Extractive vs. NLG-Based Abstractive Summarization of Evaluative Text: The Effect of Corpus Controversiality. *Proceedings of the International Natural Language Generation Conference*, 33-41. http://dx.doi.org/10.3115/1708322.1708330

[19] Lerman, K., Blair-Goldensohn, S. and McDonald, R. (2009) Sentiment Summarization: Evaluating and Learning User Preferences. *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, 514-522. http://dx.doi.org/10.3115/1609067.1609124

[20] NLTK.org. (n.d.). Retrieved 01 28, 2016. http://www.nltk.org/index.html

[21] Powers, D. (2011) Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, **2**, 37-63.

[22] Gaines, P. (2016) Accuracy, Precision, Mean and Standard Deviation. ICP Operations Guide Part 14, On 25 May. http://www.inorganicventures.com/accuracy-precision-mean-and-standard-deviation