IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.8, Issue.3, Mar 2020 E-ISSN: 2347-2693

A Survey on Digital Content Sentiment Features and Techniques

Sukhlal Sangule[1*], Sunil Phulre[2]

[1] S. V. Polytechnic College, Bhopal
[2]LNCT University, Bhopal

DOI: https://doi.org/10.26438/ijcse/v8i3.114118 | Available online at: www.ijcseonline.org

Received: 16/Feb/2020, Accepted: 17/Mar/2020, Published: 30/Mar/2020

Abstract— Sentiment analysis is one of the fastest growing research areas in computer science, making it challenging to keep track of all the activities in the area. In recent years, sentiment analysis has shifted from analyzing online product reviews to social media texts from Twitter and Facebook. Many topics beyond product reviews like stock markets, elections, disasters, medicine, software engineering, etc. extend the utilization of sentiment analysis. This paper discusses in details the various techniques to Sentiment Analysis, so class of sentiment identify accurately. Text mining pre-processing steps were also discussed for generation of features. This paper provides previous researcher work in detail. Challenges of sentiment mining were also summarized.

Index Terms—Data Mining, Opinion mining, Sentiment analysis, Text Preprocessing

I. INTRODUCTION

People are subjective creatures and conclusions are significant subsequently sentiment examination targets constructing a framework which investigations the state of mind of a person about a specific item, point or occasion communicated in a book range made in a survey, blog entry, remark or tweet. Mining assessments in a client created content is exceptionally testing however essentially extremely helpful. Estimation investigation should be possible on three levels: record level, sentence level or trait level. Report level estimation examination works by decreasing the entire archive to a solitary conclusion [1]. In any case, more often than not an archive doesn't speak to a solitary assessment. A similar report may contain diverse repudiating assessments about a similar element. Opinion investigation on sentence level arranges the assessment communicated in each sentence. The principal task is to characterize whether the sentence is emotional or objective. Sentiment mining (regularly alluded as estimation examination) is an endeavor to exploit the immense measures of client created content. It utilizes PC preparing capacity to formalize the information taken from client opinions and break down it for additional reuse [2]. In spite of the fact that there are some early works about acknowledgment of abstract writings from mid 80s and 90s, the genuine advancement in the region began with the ascent of Web 2.0. The new kinds of Internet content implemented better approaches for information the executives which, as a result, made new issues and openings emerge. In the course of the most recent decade a colossal increment of enthusiasm for the sentiment investigation inquire about is unmistakably obvious [3].
As Sentiment Analysis is a Natural Language Processing and Information Extraction task that means to acquire essayist's emotions communicated in positive or negative remarks, questions and demands, by breaking down a huge quantity of reports. As a rule, estimation examination expects to decide the mentality of a speaker or an essayist regarding some theme or the general tonality of a record. As of late, the exponential increment in the Internet utilization and trade of popular conclusion is the main impetus behind Sentiment Analysis today [4]. The Web is a gigantic store of organized and unstructured information. The investigation of this information to separate dormant general assessment and conclusion is a difficult assignment.

Rest of paper try to presents an introduction to the topic of sentiment mining. Second section gives about content pre-processing methods. Consecutively next section givens an introduction of text mining feature terms, while fourth section of this paper gives an introduction of techniques used for sentiment mining. After this related work done by other researchers were also discuss. Finally challenges present in this research area is summarized.

II. Text Pre-Processing Techniques

Preprocessing strategy assumes a significant job in content mining methods and applications. It is the initial phase in the content mining process. This work examines three key strides of preprocessing in particular, tokenization, stop words expulsion, stemming.

Tokenization: This procedure split the grouping of strings into words. It expels all the accentuations from the content information and gives expressions of content which is called tokens [5]. Fast Miner instrument gives three different ways of parting; one is default which is ordinarily utilized called non letter character where it split based on non letter of character like spaces, commas, full-stops and so on. The subsequent mode determines character, wherein you can indicate characters dependent on which sentence is part into tokens and the third one is standard articulation, wherein the regular expression is provided to split the sentence into tokens.

Stop Words Elimination: Stop words are a division of natural language. The thought process that stop-words ought to be expelled from a book is that they make the content look heavier and less significant for experts. Evacuating stop words decreases the dimensionality of term space. The most widely recognized words in content archives are articles, relational words, and professional things, and so forth that doesn't give the significance of the records. These words are treated as stop words. Model for stop words: the, in, an, a, with, and so on. Stop words are removed from documents because those words are not measured as keywords in text mining applications [6].

Stemming: This method is used to identify the root/stem of a word. For example, the words connect, connected, connecting, connections all can be stemmed to the word "connect" [7]. The purpose of this strategy is to expel different postfixes, to diminish the quantity of words, to have precisely coordinating stems, to spare time and memory space.

Part of Speech (POS) Tagging: Tagging in characteristic language preparing (NLP) alludes to any procedure that appoints certain marks to phonetic units. It signifies the task of grammatical feature labels to writings. A PC program for this purpose is known as a tagger. Grammatical form tagging incorporates the way toward allotting one of the grammatical features to the given word. For example, the english word rust for instance is either a verb or a noun. Some of dictionary are available such as Maxent Tagger from Stanford CoreNL [8].

III. Features of Text Mining

TF: Term Frequency, which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization:

$$TF(t) = \frac{Number\ of\ Times\ term\ t\ Present\ in\ Document}{Total\ Number\ of Terms\ in\ \ Document}$$

IDF: Inverse Document Frequency, which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:

$$IDF(t) = log\left(\frac{Total\ Number\ of Documents}{Number\ of\ Documents\ having\ term\ t}\right)$$

TF-IDF (Term Frequency Inverse Document Frequency) This term composed by two terms: the first computes the normalized Term Frequency (TF), The number of times a word appears in a document, divided by the total number of words in that document; the second term is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

TF-IDF = TF x IDF

IV. Different Techniques of Load Balancing

1. Lexicon-based approach

The main procedure that can be utilized for emotion / sentiment analysis is the dictionary-based strategy. It utilizes a dictionary that comprises of terms with separate emotion / sentiment scores to each term. The term can be related with a solitary word, expression or figure of speech [9]. The emotion / sentiment is characterized dependent on the nearness or nonappearance of terms in the vocabulary. The vocabulary-based methodology incorporates corpus-based methodology and word reference-based methodology that are examined further.

Dictionary-based approach: The principle thought behind the word reference based methodology is to utilize lexical databases with assessment words to remove emotion / sentiment from the report. In view of [10], a lot of seed estimation words (for example great, terrible) with their polarities is gathered by hand. Toward the start, this underlying set does not need to be huge, 30 sentiment words is sufficient [11]. Subsequent stage is to utilize the polar words to enhance a set by searching up for separate equivalent words and antonyms in a lexical database. The look-into technique is iterative. At every emphasis the calculation takes refreshed arrangement of words (extended set) and searches again until there will be no new words to incorporate. At last, a lot of emotion words can be inspected with a motivation behind erasing mistakes.

Corpus-based approach: In [12] Bing Liu shows that corpus-based methodology can be connected in two cases. First case is a recognizable proof of emotion words and their polarities in the area corpus utilizing a given arrangement of emotion / sentiment words. The subsequent case is for structure another vocabulary inside the specific area from another dictionary utilizing a space corpus. The discoveries propose that regardless of whether emotion / sentiment words are space subordinate it can happen that a similar word will have inverse direction relying upon setting. The examination directed by Hazivassiloglou and McKeown [13] is noticeable in the writing about corpus-based system. Creators proposed a technique that concentrates semantic direction of conjoined descriptive words from the corpus. The procedure depends on the use of literary corpora and seed emotion / sentiment words (descriptors). Uncommon etymological standards are connected to the corpora so as to find emotion / sentiment words with relating polarities. Creators accept that modifiers have a similar extremity on the off chance that they are joined by the combination "and". In any case, the combination "yet" is utilized for connecting descriptive

words with inverse polarities. Moreover such conjunctions as "or", "either-or", "not one or the other nor" are utilized. Some of the time these standards don't appropriate. In this way, creators likewise foresee the polarities of the conjoined descriptive words to check whether the polarities are the equivalent or not, for this reason log-direct regression model is utilized. After forecast arrange, the chart is gotten that gives interfaces between descriptive words. At that point bunching is completed on the chart to isolate descriptors into positive and negative subsets.

2. Decision tree

It is another approach to perform characterization. Decision tree [14] is a classifier that is displayed as various hierarchical decay of data space. The tree structure contains two sorts of nodes: leaf node (contains the estimation of the objective quality, for example positive or negative mark in twofold order assignment) and choice node (contains a condition on one of the properties for space division). The division of the information space is done recursively in hierarchical structure of decision tree.

3. Supervised machine learning:

These techniques accept the labeled data that are utilized for the learning procedure. Once training data is pass than obtained output is compared with desired one if class match than new data is use for training otherwise updation of weight were done. As training informational collection, marked reports must be utilized. Typically, bag of words in [15] method was utilized to speak to a report as a feature vector $d = (w1,2,\dots ,wi,\dots ,wN)$, where N is set of all the one of a kind terms in the preparation dataset and wi is weight of the I-th term. To change over training dataset to an element vector, dictionary with N exceptional words must be made from the input training dataset.

4. Unsupervised machine learning methods

Un-supervised learning methodology utilizes unlabeled datasets so as to find the structure and locate the comparable patterns from the information. This technique is generally utilized when a gathering of dependable clarified dataset is not known, yet collection of unlabeled information is simpler. It doesn't cause any troubles when new class information must be recovered. Turney [16] utilizes un-supervised approach for the comments characterization. Comments are ordered into prescribed (approval) and not suggested (disapproval) classes. The author recovers states that comprise of two words dependent on labels designs. The examples are planned so that they need to catch sentiment phrases. Each expression is a mix of descriptor/qualifier and action word/thing (by and large, 5 examples are proposed). Grammatical feature tagger is utilized to the archive so as to choose which expressions must be recovered.

V. Related Work

Fersini, Messina, and Pozzi (2016) [17] have compared the majority voting rule with other approaches, using three types of subjective signals: adjectives, emotions, emphatic expressions and expressive elongations. They re- port that adjectives are more impacting that the other considered signals, and that the average rule is able to ensure better performance than other types of rules.

Oscar Araque 2017, [18] first, this paper develops a deep learning based sentiment classifier using a word embeddings model and a linear machine learning algorithm. This classifier serves as a baseline to compare to subsequent results. Second, this paper proposes two ensemble techniques which aggregate our baseline classifier with other surface classifiers widely used in Sentiment Analysis. Third, this paper also proposes two models for combining both surface and deep features to merge information from several sources. Fourth, this paper introduces a taxonomy for classifying the different models found in the literature, as well as the ones we propose.

In [19] a novel method for extracting the hierarchical structure of Web video groups based on sentiment-aware signed network analysis is presented to realize Web video retrieval. First, the proposed method estimates latent links between Web videos by using multimodal features of contents and sentiment features obtained from texts attached to Web videos. Thus, our method enables construction of a signed network that reflects not only similarities but also positive and negative relations between topics of Web videos. Moreover, an algorithm to optimize a modularity-based measure, which can adaptively adjust the balance between positive and negative edges, was newly developed. This algorithm detects Web video groups with similar topics at multiple abstraction levels; thus, successful extraction of the hierarchical structure becomes feasible. By providing the hierarchical structure, users can obtain an overview of many Web videos and it becomes feasible to successfully retrieve the desired Web videos.

In [20] new method for the calculation of polarities and strengths of Chinese sentiment phrases is proposed in this study, which could be used to analyze semantic fuzziness of Chinese. It uses a probability value, rather than a fixed value for the polarity strengths of sentiment phrases, compared with the conventional methods. According to the polarities and strengths of those phrases, this paper proposes two multi-strategy sentiment analysis methods respectively based on SVM and NB. Particularly, in the method based on NB, this paper considers adversative conjunctions. The two methods could be used for the sentiment analysis of documents.

In [21], paper this paper focus on how to fuse textual information of Twitter messages and sentiment diffusion patterns to obtain better performance of sentiment analysis on Twitter data. To this end, this paper first analyzes sentiment diffusion by investigating a phenomenon called sentiment reversal, and find some interesting properties of sentiment reversals. Then this paper considers the inter-relationships between textual information of Twitter messages and sentiment diffusion patterns, and propose an

iterative algorithm called SentiDiff to predict sentiment polarities expressed in Twitter messages.

In [22], a Chinese sentiment analysis method based on extended dictionary is proposed. The main task of the research is to construct an extended sentiment dictionary covering fields: hotel, digital, fruit, clothing and shampoo. The extended sentiment dictionary contains the basic sentiment dictionary, some field sentiment words and polysemic sentiment words in the fields. The naive Bayesian field classifier is used to classify the field of the text in which the polysemic sentiment word is, so the sentiment polarity of the word could be distinguished.

In [23] the scalability issue that arises as the feature-set grows a novel genetic algorithm (GA)-based feature reduction technique is proposed. Hybrid sentiment analysis framework by combining ML and lexicon-based approaches in order to solve the limitations of each method. The fitness function utilizes Sentiment WordNet to evaluate feasible solutions which result in improved system scalability. Here researchers are able to reduce the feature-set size by up to 42% without compromising the accuracy.

Table 1 Comparison of Various techniques

Paper	Proposed Technique	Limitation
In [25] 2019	Genetic Algorithm based feature reduction	Fitness function need improvement
In [24] 2019	Bayesian classifier	Find sentiment of phrases in active, passive and neural class only.
In [23] 2018	Sentiment reversal Technique	Development of reverse to identify sentiment is complex and time takes.
In [22] 2018	SVM and Bayesian classifier	Find sentiment of phrases content into two class only.
In [21] 2018	Hierarchical Clustering	Classify content on the bases of content title only.
In [24] 2016	Micro blog specific sentiment lexicon	Only for Chinese blog

.

VI. Research Challenges in Sentiment Analysis

- Detection of spam and fake reviews: The web contains both genuine and spam substance. For powerful Sentiment grouping, this spam substance ought to be wiped out before handling. This should be possible by distinguishing copies, by identifying anomalies and by considering notoriety of commentator [1].
- Limitation of classification filtering: There is a restriction in order sifting while at the same time deciding most mainstream thought or idea. For better notion classification result this constraint ought to be decreased. The danger of channel bubble [11] offers superfluous input sets and it results bogus rundown of opinion.
- Asymmetry in availability of opinion mining software: The sentiment mining programming is over the top expensive and as of now reasonable just to enormous associations and government. It is past the regular citizen's desire. This ought to be accessible to all individuals, with the goal that everybody gets advantage from it.
- Incorporation of opinion with implicit and behavior data: For effective examination of conclusion, the opinion words ought to coordinate with certain information. The understood information decides the genuine conduct of notion words.
- Domain-independence: The greatest test looked by opinion mining is the space subordinate nature of sentiment words. One features set may give excellent execution in one space, simultaneously it performs exceptionally poor in some other area.
- Natural language processing overheads: The common language overhead like equivocalness, co-reference, Implicitness, derivation and so on made prevention in feeling investigation as well.
- The sentiment given on twitter is hard to grasp as it comprises of poor words, absence of capital letters, spelling botches, no appropriate accentuations, and syntactic blunders, etc.
- The sentiment of the analyst changes after some time. An exploration work is done to perceive how the state of mind of the individuals fluctuates after some time in [9]. The look into done watches destinations where the temperament of the analyst is unmistakably indicated either by browsing a given rundown of dispositions or by composing it as free content sentence.

VII. Conclusions

Opinion mining and sentiment analysis have wide range of applications. As well as there are many challenges to researches focusing in this field. This therefore has been a very active research area in recent years. With the constant advent of novel researches, a new inclusive survey is required for better understanding of current research progress. The goal of this paper is to give an in-depth introduction and present new insights toward this area. Here paper has brief previous researcher work for sentiment identification from web content. Some of the approaches gives fruitful output with content pre-processing steps. Here paper has brief features used for text content classification or analysis. In future it is desired to develop a learning model which can adaptively identify content sentiment without manual involvement.

References

[1] Kiruthika M., Sanjana Woona and Priyanka Giri(2016), "Sentiment Analysis of Twitter Data", International Journal of Innovations in Engineering and Technology(IJIET).

[2] Metin Bilgin and Izzet Fatih Senturk(2017)," Sentiment Analysis on Twitter data with Semi-Supervised Doc2Vec", UBMK International conference on computer science and engineering.

[3] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards RealTime Object Detection with Region Proposal

Networks.[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016:1-1.
[4] Lin Yue, Weitong Chen, Xue Li, Wanli Zuo, Minghao Yin. "A survey of sentiment analysis in social media". Springer-Verlag London Ltd., part of Springer Nature 2018.
[5] V. Subramaniyaswamy S. Harshaa M. Padma Janani B.S. Prabhalammbeka . "Sentiment analysis using string token classification algorithm". International Journal of Pure and Applied Mathematics 119(12):13287-13294 · January 2018.
[6] Lawrence Reeve and HyoilHan."Survey of semantic annotation platforms ACM symposium on Applied computing",. ACM, 2005, pages 1634-1638.
[7] [7]EladSegev and Regula Miesch. "A systematic procedure for detecting news biases: The case of israel in european news sites. International Journal of Communication", 2011.
[8] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, ``The Stanford CoreNLP natural language processing toolkit," in Proc. 52nd Annu. Meeting Assoc. Comput. Lin-guistics, Syst. Demonstrations, 2014, pp. 55_60.
[9] Liu, B. (2012). Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers.
[10] Hailong, Z., Wenyan, G., & Bo, J. (2014, September). Machine learning and lexicon based methods for sentiment classification: A survey. In Web Information System and Application Conference (WISA), 2014 11th (pp. 262-265)
[11] Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177).
[12] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.
[13] Hatzivassiloglou, V., & McKeown, K. R. (1997, July). Predicting the semantic orientation of adjectives. In Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics (pp. 174-181)
[14] Aggarwal, C. C., &Zhai, C. (2012). A survey of text classification algorithms. In Mining text data (pp. 163-222). Springer US.
[15] Tang, B., Kay, S., & He, H. (2016). Toward optimal feature selection in naive Bayes for text categorization. IEEE Transactions on Knowledge and Data Engineering, 28(9), 2508-2521
[16] Turney, P. D. (2002, July). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 417-424).
[17] Fersini, E. , Messina, E. , & Pozzi, F. (2016). Expressive signals in social media lan- guages to improve polarity detection. Information Processing & Management, 52 , 20–35
[18] Oscar Araque, Ignacio Corcuera-Platas, J.Fernando Sánchez-Rada, CarlosA. Iglesias. "Enhancing deep learning sentiment analysis with ensemble techniques in social applications". Expert Systems With Applications 77 (2017) 236–246.
[19] Ryosuke Harakawa, Takahiro Ogawa, And Miki Haseyama."Extracting Hierarchical Structure Of Web Video Groups Based On Sentiment-Aware Signed Network Analysis". September 19, 2017.Digital Object Identifier 10.1109/Access.2017.2741098.
[20] Ying Fang, Hai Tan, And Jun Zhang. "Multi-Strategy Sentiment Analysis Of Consumer Reviews Based On Semantic Fuzziness". Ieee Access May 2, 2018.Digital Object Identifier 10.1109/Access.2018.2820025.
[21] Lei Wang, JianweiNiu, And Shui Yu. "Sentidiff: Combining Textual Information and Sentiment Diffusion Patterns For Twitter Sentiment Analysis". Journal Of Latex Class Files, Vol. 14, No. 8, August 2018.
[22] GuixianXu ,Ziheng Yu , Haishen Yao, Fan Li, Yueting Meng, And Xu Wu. "Chinese Text Sentiment Analysis Based On Extended Sentiment Dictionary". Ieee Access April 13, 2019. Digital Object Identifier 10.1109/Access.2019.2907772
[23] Farkhund Iqbal, Jahanzeb Maqbool Hashmi, Benjamin C. M. Fung, Rabia Batool, Asad Masood Khattak, Saiqa Aleem, And Patrick C. K. Hung. "A Hybrid Framework for Sentiment Analysis Using Genetic Algorithm Based Feature Reduction". IEEE Access volume 7 February 8, 2019.
[24] Fangzhao Wu, Yongfeng Huang, Yangqiu Song, Shixia Liu," Towards building a high quality micro blog-specific Chinese sentiment lexicon", Decision Support Systems-2016.

