International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:15, No:7, 2021

1/Sigma Term Weighting Scheme for Sentiment Analysis

Hanan Alshaher, Jinsheng Xu

***Abstract*—**Large amounts of data on the web can provide valuable information. For example, product reviews help business owners measure customer satisfaction. Sentiment analysis classifies texts into two polarities: positive and negative. This paper examines movie reviews and tweets using a new term weighting scheme, called one-over-sigma (1/sigma), on benchmark datasets for sentiment classification. The proposed method aims to improve the performance of sentiment classification. The results show that 1/sigma is more accurate than the popular term weighting schemes. In order to verify if the entropy reflects the discriminating power of terms, we report a comparison of entropy values for different term weighting schemes.

***Keywords*—**Sentiment analysis, term weighting scheme, 1/sigma.

I. Introduction

SOCIAL networks, e-commerce web sites, news reports, and other web resources serve as platforms to express opinions. Opinions are subjective expressions that describe people's viewpoints, perspectives, or feelings about entities or events, and their properties. Sentiment analysis is the process of extracting and classifying sentiments, opinions, emotions, and attitudes from text. Sentiment analysis is also called opinion mining. Sentiment analysis uses natural language processing (NLP) to achieve various goals, such as measuring customer satisfaction, predicting movie sales, observing the public mood, evaluating product quality, and observing market intelligence [2], [3].

Sentiment analysis involves different tasks such as polarity classification, subjectivity detection, and summarization [4], [5]. Polarity classification determines the sentiment orientation of the text, i.e., whether it is a positive or negative opinion. Sentiment classification can be conducted at different levels such as the document [6], [7], sentence [8] or aspect level [9]. To improve polarity classification, conducting subjectivity detection is fundamental. Subjectivity detection determines whether a sentence is subjective (e.g., personal feelings or opinions) or objective (e.g., plot summaries in a movie review or some factual information). For example, [10] proposed a method to extract sentiment vocabularies by studying the dependencies among words. Finally, summarization means analyzing opinions from a large number of people to create a summary of opinions. For example, [11] created an opinion summarization for product reviews because a single opinion is not enough for decision-making. In addition, the accuracy rate and the performance of sentiment classification can be improved by using a robust term weighting scheme. Most existing research on sentiment analysis has focused on the extraction of sentiment features. Only a few studies have focused on weighting these features [1], [2]. After the extraction of the features, term weighting is a critical phase of sentiment classification; it is used to assign a value of importance to each feature. This study examined a term weighting scheme, 1/sigma, reported in [12].

This paper is structured as follows: Section II provides background information and discusses related studies. The experimental results and approach to evaluation are presented in Sections III and IV. Finally, Section VI offers a conclusion.

Hanan Alshaher is with Majmaah University and North Carolina A&T State University, Greensboro, NC 27411 USA (e-mail: halshahr@aggies.ncat.edu).

Dr. Jinsheng Xu is with the Department of Computer Science, North Carolina A&T State University, Greensboro, NC 27411 USA (e-mail: jxu@ncat.edu).

II. Related Work

Term weighting is the process of assigning weights (i.e., values) to terms. These weights measure the importance of terms to the documents. Knowing the extent to which these terms contribute to documents is fundamental for information retrieval, text classification, text clustering, and sentiment analysis.

Recently, sentiment analysis has become a popular research topic, and several studies such as [2], [14] have focused on different term weighting schemes in order to improve the performance of sentiment analysis. In [13], the authors compared the performance of sentiment analysis using supervised and unsupervised term weighting schemes, and they concluded that supervised term weight schemes outperformed unsupervised schemes. They conducted their experiment using three different datasets: Cornell Movie Review Data, the Stanford Large Movie Review Dataset, and the Multi-Domain Sentiment Dataset [22].

Nguyen et al. [14] proposed a supervised term weighting scheme called term frequency Kullback-Leibler (TF*KL) which weights each term to the ratio of its document frequency, and they used a support vector machine (SVM) for classification. The accuracy rate they achieved was 90.75%, 75.33%, and 89.35% for movie review, sentiment categorization, and multi-domain sentiment datasets, respectively. Jianqiang and Xiaolin [15] evaluated the effects of different preprocessing methods on sentiment classification, such as replacing the contractions terms (e.g., transforming "won't", "can't", and "n't" into "will not", "cannot", and "not", respectively) and removing the URLs and the stop words. To carry out their experiment, they used five Twitter datasets (Stanford Twitter Sentiment Test 140, Stanford Twitter Sentiment Gold, Sentiment Strength Twitter, Sentiment Dataset [23], and SemEval-2014 task 9). The results showed that sentiment classification was improved after expanding the acronyms and replacing the contractions terms. Moreover, removing the URLs, stop words, or numbers

had no effect on sentiment classification performance.

TABLE I
THE COMPLETE RESULTS ON STANFORD LARGE MOVIE REVIEW DATASET

Term weight schema	Accuracy	F1 score
max MI	0.8804	0.880367824
min MI	0.8816	0.881559901
1/sigma	0.9004	0.900384683
max 1/sigma	0.8948	0.894785842
min 1/sigma	0.9084	0.908372893
tf	0.8808	0.880766347
tf*IG	0.7408	0.740384615
MB25	0.8824	0.882369887
t_df	0.8976	0.897576336
CHI max	0.8564	0.856398139
CHI min	0.8564	0.856398139

TABLE II
THE COMPLETE RESULTS ON STANFORD TWITTER SENTIMENT 140 DATASET

Term weight schema	Accuracy	F1 score
tf	0.75	0.738235583
tf*idf	0.788461538	0.775978408
tf*ig	0.846153846	0.780758808
tf*1/(sig+0.001)	0.807692308	0.786718015
max MI	0.75	0.738235583
min MI	0.75	0.738235583
MB25	0.75	0.73823558
CHI max	0.865	0.8408271
CHI min	0.865	0.8408271

The Cornell Movie Review Dataset has been used in many studies, such as [16] and [17]. Prabowo and Thelwall [16] using SVM as a classifier reported an accuracy rate of 87.30% while Boiy and Moens [17] using maximum entropy (MaxEnt) as a classifier reported an accuracy rate of 87.40%. Pang and Lee [6] combined Naive Bayes (NB) and SVM classifiers to obtain an accuracy rate of 87.2%. The Twitter dataset was used in [18], and the results showed that unigram extraction of features resulted in an accuracy rate of 81.3%, 80.5%, and 82.2% for the NB, MaxEnt, and SVM classifiers, respectively, while combining unigram and bigram extraction resulted in 82.7%, 83.0%, and 81.6%, respectively.

From the previous works, it is clear that a term weighting scheme plays a significant role in improving the sentiment classification accuracy. This paper evaluates a term weighting scheme, one-over-sigma (1/sigma), which was proposed in [12] on two different sentiment datasets.

III. EXPERIMENTAL EVALUATION

A. Dataset

We examined the 1/sigma term weighting scheme on different types of datasets for sentiment classification since it delivered promising results in a previous study for authorship identification [12]; the results showed that the proposed scheme outperformed other popular term weighting schemes for authorship identification. Two different sentiment datasets, which served as a benchmark for sentiment classification, were used in our experiment. A brief description of each dataset that was used in the present study is presented below.

TABLE III
THE TOP 20 TERMS WITH HIGHEST TERM WEIGHT FOR TF SCHEME FROM STANFORD TWITTER SENTIMENT 140 DATASET

Term	**TF**	Probability	Entropy
the	**45.503**	0.025618	0.135433
I	**34.54902**	0.019451	0.110558
http	**34.07002**	0.019181	0.109412
to	**33.54328**	0.018884	0.108145
a	**29.28055**	0.016485	0.097634
is	**28.31851**	0.015943	0.095194
and	**21.85333**	0.012303	0.078061
at	**16.53144**	0.009307	0.062799
s	**16.00024**	0.009008	0.061205
bit	**15.6423**	0.008806	0.060123
for	**15.44114**	0.008693	0.059512
ly	**15.21927**	0.008568	0.058836
my	**14.45076**	0.008136	0.056473
of	**13.32044**	0.007499	0.052937
in	**13.26394**	0.007467	0.052759
it	**13.00495**	0.007322	0.051937
com	**12.41822**	0.006991	0.050059
with	**12.18614**	0.006861	0.04931
i	**9.503586**	0.00535	0.040375

1. Stanford Large Movie Review Dataset v1.0

Stanford Large Movie Review Dataset [19] contains 50k movie reviews along with their associated sentiment labels. The original dataset is divided into training and testing folders. Each folder contains 25k reviews divided into 12,500 positive reviews and 12,500 negative reviews. Because the Stanford Movie Review Dataset contains a large number of terms (i.e., features), we only used the training folder; and then we further divided it into two parts: training and testing. Each folder contains positive and negative reviews. Since the training folder contains 25k reviews, we used "0" as a filter to split the review files into 22,500 reviews for training and 2500 reviews for testing. Consequently, we obtained a balanced distribution of labels (i.e., 11,250 positive reviews for training and 1250 positive reviews for testing, and the same numbers for the negative reviews).

2. Stanford Twitter Sentiment

Test 140, the Stanford Twitter sentiment test (i.e., Sentiment140) dataset, is a computer-separated values (CSV) file [18]. The CSV file contains six fields: the label of the tweet (e.g., 0 = negative, 2 = neutral, and 4 = positive), the serial number of the tweet, the date of the tweet, and the query (for tweets that do not have a query, the value is NO QUERY), the user that tweeted, and the text of the tweet. The dataset contains 177 negative, 182 positive, and 139 neutral tweets.

B. Preprocessing

In the present study, tokenization is used to break the sentences into words. Even though stop words do not contribute to sentiment analysis, as tested in [15], during the preprocessing step no attempt was made to alter the dataset,

such as dropping stop words.

TABLE IV
THE TOP 20 TERMS WITH HIGHEST TERM WEIGHT FOR 1/SIGMA AND TF-IDF SCHEME FROM STANFORD TWITTER SENTIMENT 140 DATASET

Term	**TF*1/sigma**	probability	entropy	Term	**TF*IDF**	probability	entropy
the	**347.9267**	0.004705	0.036374	the	**53.40448**	0.007131	0.050856
http	**330.2581**	0.004466	0.034863	I	**50.63739**	0.006762	0.04874
to	**281.3408**	0.003804	0.030579	to	**48.87288**	0.006526	0.047376
a	**279.5606**	0.00378	0.03042	http	**45.75828**	0.00611	0.044937
I	**276.4099**	0.003738	0.030138	is	**45.46338**	0.006071	0.044704
is	**258.0256**	0.003489	0.02848	a	**43.95216**	0.005869	0.043504
and	**226.7399**	0.003066	0.025599	and	**39.15591**	0.005228	0.039629
s	**193.3025**	0.002614	0.022425	at	**34.7115**	0.004635	0.035936
bit	**189.81**	0.002567	0.022088	bit	**33.1031**	0.00442	0.034574
ly	**185.9972**	0.002515	0.021718	ly	**32.72382**	0.00437	0.03425
at	**185.8234**	0.002513	0.021701	for	**32.6774**	0.004363	0.03421
for	**184.2795**	0.002492	0.02155	s	**32.57989**	0.00435	0.034127
of	**180.0568**	0.002435	0.021138	my	**32.10404**	0.004287	0.03372
it	**173.7926**	0.00235	0.020523	in	**29.71532**	0.003968	0.031653
my	**172.4306**	0.002332	0.020388	with	**29.02687**	0.003876	0.031051
in	**170.6599**	0.002308	0.020213	it	**28.89201**	0.003858	0.030933
com	**166.2822**	0.002248	0.019779	of	**28.87269**	0.003855	0.030916
with	**155.2723**	0.0021	0.018677	com	**28.79512**	0.003845	0.030848
on	**138.3193**	0.00187	0.01695	i	**25.5099**	0.003406	0.027924

C. Methodology

Term frequency (TF) is the classic method used to describe the content of a document. Several functions are used to compute the importance of the term for the document, such as information gain (IG), mutual information (MI), chi-squared statistic (x2 or CHI), and inverse document frequency (IDF), or best match 25 (BM25). These functions provide a statistical distribution of the term among the documents. The weight of the term is measured by multiplying the term frequency by one of these functions. Seven term weighting schemes were examined in our experiment. According to [13], some of these term weighting schemes are unsupervised (e.g., TF and BM25), and others are supervised (e.g., TF*IG and TF*CHI). Moreover, since the SVM classifier had a better performance than the other base classifiers [12], and it is widely used in the field of sentiment analysis [13], we used SVM as the classifier to investigate the performance of the seven different term weighting schemes. We used accuracy and the f1 score to measure the sentiment analysis performance.

IV. RESULTS

This section presents a comparison of the results in [13] and [15] and the results of the proposed term weighting scheme (i.e., 1/sigma). Table I shows the accuracy rate and f1 score for seven different term weighting schemes using the Stanford Large Movie Review Dataset. The 1/sigma scheme was found to have the highest accuracy rate at 90.04%, followed by the TF-IDF at 89.76%. The results for the other term weighting schemes are very similar to the accuracy rates reported in [13], which were 88.008%, 87.771%, and 87.096% for MI, TF*WFO when $\lambda = 0.1$ and MB25, respectively. Moreover, Table II shows the accuracy rate and f1 score for seven different term weighting schemes using the Stanford Twitter Sentiment Test 140 dataset. The TF*IG scheme had the highest accuracy rate of 84.61%, followed by 1/sigma, which achieved an accuracy rate of 80.76%. The 1/sigma schema obtained the highest f1 score, which was 78.67%, followed by TF*IG at 78.07%. Jianqiang and Xiaolin [15] compared the Stanford Twitter dataset results to the baseline results and reported that accuracy was improved by 0.58% using the SVM classifier.

V. DISCUSSION

In [1], the authors investigated the limitations of TF and TF-IDF term weighting schemes in the context of text classification. They studied the relationship between a term's weight and its entropy. They stated that a term with a high weight has a smaller entropy. Thus, term classification depends on the entropy, and the classifier selects the term with the largest entropy. Furthermore, the principle of maximum entropy states that the probability distribution that best represents the current state of knowledge is the one with largest entropy [20]. Tables III and IV show the top 20 terms with the highest term weight for three different term weighting schemes from the Twitter dataset. The order of the terms is different for each term weight, since it is based on the term's weight. According to our observation, 1/sigma had the smallest entropy in comparison to the other term weighting schemes for the terms that have the highest weight. Since TF-IDF outperformed TF because of the terms' entropy, 1/sigma outperformed TF-IDF for the same reason. Moreover, Table V shows the sum of entropy for each term weighting scheme on two different datasets; it is clear that 1/sigma has the highest entropy in comparison to the other term weighting schemes. However, [21] claimed that TF-IDF does not reflect the

distribution of terms in the text, while 1/sigma measures how terms are spread out over documents. Since one-over-sigma has the largest entropy in Table V, based on the principle of maximum entropy the probability distribution of 1/sigma represents the best current state of knowledge. In short, 1/sigma outperformed other term weighting schemes in text classification [12] and sentiment classification.

TABLE V
THE SUM OF ENTROPY FOR DIFFERENT TERM WEIGHT SCHEMES

Dataset	Term weight scheme	Sum of terms' weight	Entropy
Stanford Large Movie Review Dataset	TF	207321.3659	10.27009208
	1/sigma	21758836.71	14.45711238
	TF-IDF	483673.6987	13.69416447
Stanford Twitter Dataset 140	TF	1776.242	9.854354
	1/sigma	73955.45	11.1252
	TF-IDF	7488.98	10.71185

VI. CONCLUSION

A few studies on sentiment analysis have focused on weighting sentiment features. Each feature (i.e., term) in a textual document makes a differently important contribution to the text; hence, a term weighting scheme measures weight (wi) for each term (ti) in a document (d). The weight value (wi) (usually between 0 and 1) represents how much the (ti) term contributes to document (d). Term weighting improves the performance of text classification, for example as seen in [12]. This study examined a term weighting scheme, 1/sigma, reported in [12]. The results show that 1/sigma outperformed seven different term weighting schemes. Moreover, the entropy comparison between 1/sigma and two popular term weighting schemes (i.e., TF and TF-IDF) showed that 1/sigma has the smallest entropy values for terms with the largest term weight, which improves the accuracy of the sentiment classification performance.

For future work, we will theoretically analyze why the proposed term weighting scheme (i.e., 1/sigma) performed better under the SVM classification compared to the baseline term weight schemes TF and TF-IDF.

REFERENCES

[1] Wang, T., Cai, Y., Leung, H.F., Cai, Z. and Min, H., 2015, November. *Entropy-based term weighting schemes for text categorization in VSM*, In 2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 325-332). IEEE.

[2] Zhang, P.,Wang, Y.,Wang, J., Zeng, X. and Wang, Y., 2017, March. *An improved term weighting scheme for sentiment classification*, In 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC) (pp. 462-466). IEEE.

[3] Ravi, K. and Ravi, V., 2015. *A survey on opinion mining and sentiment analysis: tasks, approaches and applications*, Knowledge-Based Systems, 89, pp.14-46.

[4] Ismail, H., Harous, S. and Belkhouche, B., 2016. *A Comparative Analysis of Machine Learning Classifiers for Twitter Sentiment Analysis*, Res. Comput. Sci., 110, pp.71-83.

[5] Liu, B., 2012. *Sentiment analysis and opinion mining. Synthesis lectures on human language technologies,* 5(1), pp.1-167.

[6] [6] Pang, B. and Lee, L., 2004. *A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts*, arXiv preprint cs/0409058.

[7] Untawale, T.M. and Choudhari, G., 2019, March. *Implementation of Sentiment Classification of Movie Reviews by Supervised Machine Learning Approaches*, In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1197-1200). IEEE.

[8] Tang, D., Qin, B., Wei, F., Dong, L., Liu, T. and Zhou, M., 2015. *A joint segmentation and classification framework for sentence level sentiment classification*, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(11), pp.1750-1761.

[9] Zhou, J., Huang, J.X., Chen, Q., Hu, Q.V., Wang, T. and He, L., 2019. *Deep learning for aspect-level sentiment classification: Survey, vision, and challenges*, IEEE Access, 7, pp.78454-78483.

[10] Bai, X., 2011. *Predicting consumer sentiments from online text.* Decision Support Systems, 50(4), pp.732-742.

[11] Hu, M. and 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] Alshaher, H. and Xu, J., 2020, March. *A New Term Weight Scheme and Ensemble Technique for Authorship Identification*, In Proceedings of the 2020 the 4th International Conference on Compute and Data Analysis (pp. 123-130).

[13] Deng, Z.H., Luo, K.H. and Yu, H.L., 2014. *A study of supervised term weighting scheme for sentiment analysis*, Expert Systems with Applications, 41(7), pp.3506-3513.

[14] Nguyen, T.T., Chang, K. and Hui, S.C., 2011, July. *Supervised term weighting for sentiment analysis*, In Proceedings of 2011 IEEE International Conference on Intelligence and Security Informatics (pp. 89-94). IEEE.

[15] Jianqiang, Z. and Xiaolin, G., 2017. *Comparison research on text preprocessing methods on twitter sentiment analysis*, IEEE Access, 5, pp.2870-2879.

[16] Prabowo, R. and Thelwall, M., 2009. *Sentiment analysis: A combined approach*, Journal of Informetrics, 3(2), pp.143-157.

[17] Boiy, E. and Moens, M.F., 2009. *A machine learning approach to sentiment analysis in multilingual Web texts*, Information retrieval, 12(5), pp.526-558.

[18] Go, A., Bhayani, R. and Huang, L., 2009. *Twitter sentiment classification using distant supervision*, CS224N project report, Stanford, 1(12), p.2009.

[19] Maas, A., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y. and Potts, C., 2011, June. *Learning word vectors for sentiment analysis*, In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies (pp. 142- 150).

[20] Kantor, P.B. and Lee, J.J., 1986, September. *The maximum entropy principle in information retrieval*, In Proceedings of the 9th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 269-274).

[21] Kuang, Qiaoyan, and Xiaoming Xu. *Improvement and application of TF• IDF method based on text classification*, 2010 International Conference on Internet Technology and Applications. IEEE, 2010.

[22] Dredze, Mark. *Multi-Domain Sentiment Dataset (Version 2.0)*. Johns Hopkins University, 23 Mar. 2009, https://www.cs.jhu.edu/~mdredze/datasets/sentiment/.

[23] Narr, Sascha, Michael Hulfenhaus, and Sahin Albayrak. "Language-independent twitter sentiment analysis." *Knowledge discovery and machine learning (KDML), LWA* (2012): 12-14.