SN Computer Science (2021) 2:414
https://doi.org/10.1007/s42979-021-00807-1

ORIGINAL RESEARCH

Specialists, Scientists, and Sentiments: Word2Vec and Doc2Vec in Analysis of Scientific and Medical Texts

Qufei Chen[1] · **Marina Sokolova**[2]

Received: 23 December 2020 / Accepted: 2 August 2021 / Published online: 15 August 2021

Abstract

Analyze performance of unsupervised embedding algorithms in sentiment analysis of knowledge-rich data sets. We apply state-of-the-art embedding algorithms Word2Vec and Doc2Vec as the learning techniques. The algorithms build word and document embeddings in an unsupervised manner. To assess the algorithms' performance, we define sentiment metrics and use a semantic lexicon SentiWordNet (SWN) to establish the benchmark measures. Our empirical results are obtained on the Obesity data set from i2b2 clinical discharge summaries and the Reuters Science dataset. We use the Welch's test to analyze the obtained sentiment evaluation. On the Obesity data, the Welch's test found significant difference between the SWN evaluation of the most positive and most negative texts. On the same data, the Word2Vec results support the SWN results, whereas the Doc2Vec results partially correspond to the Word2Vec and the SWN results. On the Reuters data, the Welch's test did not find significant difference between the SWN evaluation of the most positive and most negative texts. On the same data, Word2Vec and Doc2Vec results only in part correspond to the SWN results. In unsupervised sentiment analysis of medical and scientific texts, the Word2Vec sentiment analysis has been more consistent with the SentiWordNet sentiment assessment than the Doc2Vec sentiment analysis. The Welch's test of the SentiWordNet results has been a strong indicator of future correspondence between Word2Vec and SentiWordNet results.

Keywords Unsupervised sentiment analysis · Word2Vec · Doc2Vec · Clinical discharge summaries · Reuters science data

Introduction

Unsupervised sentiment analysis, a field of natural language processing and machine learning, works with data that does not have sentiment labels assigned to it. Unsupervised sentiment analysis is a growing-in-importance part of sentiment analysis, which is by itself a state-of-art study of unstructured text. Sentiment analysis, in both supervised and unsupervised models, mostly works with texts retrieved from social media, e.g., texts harvested from Twitter, Facebook, Reddit [21] or the databases of product and movie reviews [44]. In those data sets, expressed sentiments can be reliably assigned either through *stars* in customer-written reviews, emoticons, or *hashtags* and *emoji* in Twitter [26]. This sentiment metadata significantly contributes to analysis of sentiments [27].

At the same time, commonly used sentiments analysis methods do not apply to texts originating from knowledge-rich domains as science, jurisprudence, and social science. Articles published in scientific and medical journals [16, 36], popular science publications from main-stream media, documents issued by organizations [41] remains understudied by sentiment analysis, especially by unsupervised sentiment analysis. Medical and scientific texts where the author's positions may not be explicitly revealed present the highest challenge for analysis of sentiments. While medical and scientific texts appear to be objective, they often subtly express the authors' sentiments. It is important to identify sentiments in these seemingly objective texts, as the presence of unconscious bias, both positive and negative, can be extremely harmful [5].

Medical and scientific documents and professional texts are customarily written with a high volume of

✉ Marina Sokolova
sokolova@uottawa.ca

Qufei Chen
qchen037@uottawa.ca

[1] School of EECS, University of Ottawa, Ottawa, ON, Canada

[2] IBDA@Dalhousie University, University of Ottawa, SEPH, 600 Peter Morand Cres, Ottawa, ON K1G 5Z3, Canada

domain-related terms. For those texts, manual assignment of sentiments is often impractical: (a) expert annotations are prohibitively expensive; (b) sentiment labeling has a lower inter-annotator agreement than other contrived texts; for example, on the same type of texts, inter-annotator agreement has Fleiss kappa = 0.46 on medical-focused texts [3], whereas texts focused on economics and geography reach Fleiss kappa's 0.71 (economics) and 0.61 (geography) [7]. In medical subdomains, many sentiment analysis resources are built from and applied to user-written online texts [22, 33], thus making the resources less reliable in the analysis of professional and formally written medical text. So far, there are few lexical resources dedicated to analysis of scientific texts. Those that exist may focus on topics other than sentiments, e.g., subject matters of planetary science [42] or specific document constructions, e.g., Ph.D. theses [4].

In the current work, we present a novel approach for unsupervised sentiment analysis of medical and scientific texts. In this approach, we use unsupervised machine learning algorithms and engage sentiment similarity between the data sentiment evaluation through a general linguistic source and the methods' results. For the methods, we make use of two state-of-art text analytics techniques, Word2Vec and Doc2Vec. Word2Vec and Doc2Vec are unsupervised embedding algorithms comprised of shallow two-layer neural networks, successfully used in various fields of text analytics. The algorithms, esp. Word2Vec, are popular tools of sentiment analysis. However, in many studies Word2Vec and Doc2Vec are applied on user-written texts harvested from various online sources [17, 38]. We, however, work with medical and scientific texts that contain significantly fewer sentiment terms than texts posted online. To evaluate the results of the sentiment analysis of the medical and scientific texts, we use SentiWordNet (SWN), a lexical resource purposefully designed for sentiment analysis and opinion mining. We use SWN to compute the benchmark sentiment values for the data sets. We apply our approach to two large, highly granular data sets: the i2b2b Obesity NLP Challenge (medical texts) and four Science subgroups from Reuters 20 Newsgroups (scientific texts). We use Welch's *t* test to generalize on the benchmark results. Comparison with the benchmark results show that Word2vec exhibits a reliable performance in sentiment analysis of the text. We have reported preliminary results in [5, 6].

A Quick Overview of Sentiment Analysis

Sentiment Analysis

Sentiment analysis represents a computationally rich process of identifying and evaluating various aspects of subjectivity in text. Sentiment classification, polarity analysis, opinion mining, emotion and affect recognition, with a subfield of mental state recognition, and subjectivity analysis belong to the field of sentiment analysis. Sentiment analysis combines natural language processing (NLP) and machine learning methods. The results depend on the tools' ability to extract and represent lexical sentiment representation from a singular text and ability to find and generalize sentiments' lexical representations extracted from a large volume of independently written tests.

The NLP process uses linguistic knowledge of emotions, opinions, affect, and appraisal and their textual expressions [35]. To extract that knowledge from a significant volume of text, far beyond human processing capacities, the research field relies on computational methods, especially on machine learning (ibid). Machine learning tasks are usually dichotomized along supervised (i.e., data labels are known), unsupervised (i.e., data labels are not known), and semi-supervised learning (i.e., labels are known for a part of the data). The same dichotomy applies to sentiment analysis: if the text sentiment labels are known, the study belongs to supervised analysis; if the text sentiment labels are not known, the study belongs to unsupervised analysis. Unsupervised sentiment analysis is often used as an exploratory first step to identify dominant emotions in the data set. We note that in unsupervised sentiment analysis texts can be assigned with other labels, e.g., topic or subject matter.

Significant difference between supervised analysis and unsupervised analysis lies in evaluation of the results. In supervised analysis, the access to the data labels allows for a direct calculation of correct and incorrect results output by the methods. In unsupervised analysis, a reliable assessment of the results remains a research challenge.

At present time, sentiment analysis mostly deals with subjective, user-written text, often gathered from social media and social networks. Considerably less attention has been paid to study sentiments in texts that are presumed to be objective, e.g., scientific texts. In such texts, the authors do not express sentiments directly. However, those texts do convey sentiments, often in a rather subtle or even unintentional presence, underlined by the word choice and sentence structure. This results in the number of sentiment terms expressed in the text to be rather low (5–11% in clinical narratives) [10], consequently resulting in a lower accuracy of sentiment analysis [11]. In addition, scientific texts, especially medical and health care texts, use a formal language and often contain professional terms and expressions not found elsewhere. Such texts are descriptive rather than opinionated: it is more likely to find a sentence that reads "the patient presented with chest pain" rather than "the patient is doing really badly".

Consequences of detecting subjectivity in medical and scientific text are also quite significant. Sentiments found in a published text have a profound effect on social behavior

[25]. It is important to be able to detect and address the author's attitude, be it an article on conservation biology [20] or opinion about a cited work [4, 48]. Identification of negative attitudes, and possible negative biases can serve to ensure that health-care patients diagnosed with certain conditions [5] or technological advances are not discriminated against due to the author's subjectivity. Equally dangerous can be an undetected bias towards the object of discussion; that bias can promote barely tested methods and applications, and thus endanger the general public.

Lexical resources and dictionaries of affective and emotional language assist in extraction of sentiment expressions from text. Those resources can be (a) general, e.g., SentiWordNet [14], (b) domain-specific, e.g., a lexicon of textual depression identifiers [21], (c) data-based, e.g., built from the COVID-19 headlines published by the mainstream media [2], (d) platform-specific, e.g., HealthAffect, a lexicon built from texts posted on a medical forum [33]. General lexical resources have advantage of comprehensive language representation. They are data- and domain-independent and can be used in a variety of sentiment analysis tasks and applied to diverse data sets. At the same time, general linguistic resources may underperform when used in specific, topical studies [12].

Related Work

Prohibitive costs and a high subjectivity of manual annotations of knowledge-rich texts make a supervised sentiment analysis unfeasible as a supervised analysis requires a reliably labeled training data. Unfortunately, unsupervised sentiment analysis still underserves such texts. We list here the most relevant work that uses both supervised and unsupervised sentiment analysis.

In our study, Word2Vec exhibited a better performance than Doc2Vec on the Obesity data subsets. A similar outcome was observed in a supervised sentiment analysis of hospital feedback data [46]. As the texts were labeled, the authors were able to estimate accuracy of the algorithm performance. They reported that Word2Vec—CBOW achieved a better accuracy than a few Doc2Vec models. Similar results were reported in [39]. The study compared Word2Vec and Doc2Vec performance in the supervised learning of the text categories. The algorithms were applied on the Reuters 21578 data. In the multi-class text classification, Word2Vec outperformed Doc2Vec. We have shown that in unsupervised sentiment analysis Word2Vec outperforms Doc2Vec on the Obesity data set. On the Reuters Science four subsets, the results are less conclusive.

Implicit and explicit judgments expressed by radiologists and clinicians were studied in [12]. The authors counted the number of positive and negative occurrences in those texts, and then compared the results with manual annotations by clinical experts. However, the accuracy of the proposed method was only 42.0% on the nurse letters and 44.6% on the radiology reports respectively. The authors concluded that the simple count method is not well suited to analyze sentiment in clinical narratives. Word2Vec has been employed in variety of sentiment analysis studies, including polarity classification in identification of adverse drug reaction [47]. We, on the other hand, apply advanced Word2Vec and Doc2Vec methods to analyze sentiments in medical and scientific documents.

Unsupervised sentiment analysis of clinical narratives has been reported in [49]. The authors used an aspect-based information extraction (i.e., they extracted nouns and corresponding adjectives) as a supervision signal. Such signals were supplemented with several hand-crafted rules, e.g., verb inclusion, clause identification, manual identification of implicit relationships between adjectives and nouns, and manual assignment of adjectives to the target aspects. Our sentiment analysis uses an enhanced sentiment lexicon, SentiWordNet, that can be used in an "as is" setting, thus transportable to analysis of various data sets.

COVID-19 and the ensuing pandemic had further prompted applications of sentiment analysis to health-related texts. Such studies were mostly performed on user-written messages posted online, e.g. [45]. We, however, work with medical and scientific texts written by professionals.

Transformer-based methods, especially BERT, have been successfully applied to a health-related sentiment analysis [37, 43]. Those studies were performed in supervised sentiments analysis settings, when the methods were trained and tested on labelled data sets. We, on the other hand, work with unsupervised sentiment analysis.

Relevant Computational Methods, Lexical Resources, and Evaluation

Computational Methods

We use two unsupervised computational algorithms that build distributional vector representations of words and documents. Embedding algorithms derive all the necessary information from the input data, thus skip the pre-training step required by many other algorithms, including popular transformer-based methods [37, 43] (note that the pre-training step requires access to a large volume of labelled and/or annotated data). Embedding algorithms' self-sufficiency enables applications on novel tasks where pre-training is not feasible.

Word2Vec refers to a family of unsupervised shallow two-layer neural network models [23]. We have used the Word2Vec implementation from Generate Similar (Gensim) [32]. Gensim is an open source python library for semantic tools

[31]. The tools exhibit a high clarity, efficiency, and scalability. We have used the default setting provided with the source. Word embeddings are the numeric representation of words in the form of vectors. Word2Vec produces word embeddings based on the contextual semantics of words in a text (based on the context that the word occurs in). Words with similar linguistic contexts are mathematically grouped together in a vector space, which preserves the semantic relationship between words. Word2Vec can then use the embedding to produce predictions on a word's meaning. There are two model architectures of Word2Vec that can be used:

- Skip-gram model.
- Continuous bag-of-words model (CBOW).

The skip-gram model takes in a word as input and aims to predict a target context, while the continuous bag of words method takes in a context as input and aims to predict a specific word [8]. Since we would like to learn the contexts of the words in the data sets, we further employ the continuous bag-of-words model.

Doc2Vec is an extension of Word2Vec that is applied to a document as a whole instead of individual words [19]. It aims to create a numerical representation of a document rather than a word. Doc2Vec operates on the logic that the meaning of a word also depends on the document that it occurs in. The vectors generated by Doc2Vec can be used for finding similarities between documents. Doc2Vec has been used in sentiment analysis, albeit not as often as Word2Vec. We have used the Doc2Vec implementation from Gensim, with the default hyperparameters. Doc2Vec applications include sentiment gauging from advertisement notes posted on an advertisement platform [24].

Lexical Resources

We opted for *SentiWordNet* (SWN) as an independent knowledge source. The resource is an extension of WordNet,[1] a lexical database for Queen's English, i.e., standard, grammatically correct, and coherent expression in the English language. It groups words together into sets of synonyms (called synsets). SentiWordNet works on top of that by assigning each synset in WordNet three sentiment scores: a positive score, a negative score, and an objective score: pos_w, neg_w, and $\text{obj}_w = 1 - \left(\text{pos}_w + \text{neg}_w\right) = 1 - \text{sub}_w$. The positive score and the negative score make up the subjective portion of the word (w), and the objective score is calculated by taking one minus the sum of the positive and negative scores (in other words, the objective scores is equal to one minus the total subjective component of the word).

SentiWordNet has proven to be a reliable sentiment knowledge base. It performed exceedingly well in several sentiment analysis competitions. It obtained the best performing resource when sentiment analysis was conducted on the SemEval2013 dataset [28]: SentiWordNet's best accuracy was 58.99%, whereas the 2nd best accuracy 58.25% was obtained by MPQA. It again obtained the best accuracy on the Stanford Twitter Sentiment dataset [15]: SentiWordNet's performance resulted in 72.42% accuracy; the 2nd best accuracy of 70.75% was again obtained by MPQA (Musto et al. 2014).

We then calculate the cosine similarity (Eq. 1), where $V(D)$ is a vector representing a dataset D and $V(L)$ is a vector representing a sentiment list L. Cosine similarity measures the difference between vectors based on the cosine of the angle between the vectors while normalizing for length. This results in cosine similarity being a measure for the difference in orientation of the vectors in the vector space rather than a measure of magnitude.

$$\text{Cosine similarity, } \text{sim}(D, L) = \frac{\overrightarrow{V(D)}\overrightarrow{V(L)}}{|\overrightarrow{V(D)}||\overrightarrow{V(L)}|}. \tag{1}$$

A high cosine similarity between a dataset and a list indicates that the context of the dataset is similar to the context of the list, thus indicating prevalence of a positive or a negative sentiment in the set.

Evaluation

Evaluation of results is a major challenge in unsupervised sentiment analysis, as the results are obtained on a large number of unlabeled texts. If the researchers use manual annotation to confirm the obtained results, they face the same issues of reliability and affordability that we have listed in "Introduction". To overcome the evaluation challenges, we use SWN as the benchmark for assessment of our sentiment models. To access the models' efficacy in sentiment representation, we compute the sentiment agreement (Cagan et al. 2017) by calculating the cosine similarity between the model's outputs representing sentiments in the data.

Additionally, we define SWN sentiment scores Sc_T and SWN sentiment ratio R_T as the sentiment benchmarks for a data set (T). We define the positive score $\text{Sc}_{\text{pos}(T)} = \frac{\sum_{t \in N_{\text{pos}}} \text{pos}_t}{N}$, the negative score $\text{Sc}_{\text{neg}(T)} = \frac{\sum_{t \in N_{\text{neg}}} \text{neg}_t}{N}$, the objective score $\text{Sc}_{\text{obj}(T)} = \frac{\sum_{t \in N_{\text{obj}}} \text{obj}_t}{N}$, and positive ratio $R_{\text{pos}} = \frac{N_{\text{pos}}}{N}$, negative ratio $R_{\text{neg}} = \frac{N_{\text{neg}}}{N}$, where t is a term, N is the total number of terms in the dataset, N_{pos} is the number of positive terms, N_{neg} is the number of negative terms, N_{obj} is the number of objective terms in the dataset, pos_t is

[1] https://wordnet.princeton.edu.

Table 1 The data sets' stylometrics parameters ($K = 1000$)

Data	# of texts	# of subsets	Tokens	Types	Occurrence of rare words		Token match with SWN vocabulary	
					Hapaxlegomena	Dislegomena		
Obesity	1237	12	804 K	32.1 K	19.1 K	2.8 K	260.5 K	32.04%
Science	3949	4	460 K	36.8 K	17.3 K	5.4 K	304.0 K	66.10%

the positive SWN score of a term, neg_t is the negative SWN score of a term, obj_t is the objective SWN score of a term.

We employ Welch's test for further statistical generalization of the obtained results (Eq. 2).

$$\text{Welch's } t \text{ test, } t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}, \tag{2}$$

where $\overline{X}_j$, s_j, and N_j are the jth sample mean, standard deviation, and its size. The Welch's t test considered a more reliant choice for psychology studies when the assumptions of homogeneity of variance is not met [9]. It performs better than a more popular Student's t test when the samples being compared have unequal variances and sample sizes, and correspond to the Student's t test when the variances and sample sizes are equal [18].

Data Sets

We have used two data sets in our empirical study: the i2b2 Obesity data and the Reuters Science. Table 1 reports their descriptive statistics. Number of texts indicates granularity of the data set. Tokens mean the number of all words in the data sets, whereas types mean the number of different words in the data set. For example, in "discharge summaries belonging to both set A and set B", "set" is counted as two tokens and one type. Hapax legomena equals to the number of words occurring once in the data; dis legomena signifies the number of words occurring twice in the data. Note that occurrences of rare words are considerably lower than in social network texts, thus giving a larger weight to the content words [34].

1. Obesity NLP Challenge Dataset[2] is a part of the deidentified clinical discharge summaries built for an i2b2 competition. This dataset was created for information extraction of obesity and its common comorbidities (hereinafter referred to as diseases) [40]. The obesity dataset consists of 1237 de-identified clinical discharge summaries written by physicians and nurses, as well as a set of annotations that specifies for each discharge summary the occurrence of any number of diseases. We separated the discharge summaries into subsets that correspond to 17 diseases. For each discharge summary, manual annotations specify occurrence of number of diseases. The annotations split the occurrence into four classes: present, absent, questionable, and unknown. We used annotation *Present* to separate the set into subsets.

Some diseases have not been sufficiently represented for Word2Vec and Doc2Vec applications, containing too few discharge summaries, e.g., 25 summaries. For the purposes of this study, we have chosen data sets corresponding to three diseases: Hypertension (the most prevalent disease), Diabetes (the second most prevalent), and Obesity (known for invoking a negative stigma and the reason this dataset was chosen). A set of discharge summaries relating to one of these diseases will be referred to as a subset.

We observed that many discharge summaries are associated with more than one disease or condition (e.g., obesity and diabetes). The disease combinations allow for a finer-grade sentiment analysis. Therefore, in addition to the discharge summaries for the three chosen diseases (i.e., hypertension, diabetes, and obesity), we also work with the set complements (discharge summaries belonging to set A but not set B) as well as their intersections (discharge summaries belonging to both set A and set B) as additional subsets. The twelve subsets are reported in Table 2. For brevity, we assign numerical IDs for the data sets.

Table 2 The i2b2 data subsets

Annotated disease	Number of summaries	Digital ID
Obesity	443	1
Hypertension	816	2
Diabetes	737	3
Obesity and hypertension	332	4
Obesity and not hypertension	111	5
Hypertension and not obesity	484	6
Obesity and diabetes	258	7
Obesity and not diabetes	185	8
Diabetes and not obesity	479	9
Hypertension and diabetes	595	10
Hypertension and not diabetes	221	11
Diabetes and not hypertension	142	12

[2] https://www.i2b2.org/NLP/DataSets/Main.php.

Table 3 The Reuters20 data subsets

Categories	Number of texts
Crypto	1000
Space	1000
Electronics	1000
Med	1000

We used NLTK[3] to remove all the formatting text and special characters from each discharge summary, then tokenized the remaining text. Stop words were also removed using the NLTK's English stop words list.[4] To minimize mismatch with SWN vocabulary, we lemmatized the data.

2. Reuters 20 Newsgroups[5] is a collection of approximately 20,000 documents published by Reuters. The texts are sorted into 20 different categories by the topics they cover, such as space, politics, and religion. Every category has 1000 entries except Christian Religion (which has 997). We work with texts assigned to the Science group of categories: cryptography, electronics, medicine, and space. We used each category as a distinct subset, and as such we have four subsets related to Science (Table 3).

The Reuters20 is a benchmark data used in natural language processing and text data mining research. We used the data "as is", without additional pre-processing. In this study, we treat the Reuters data as a control data.

Building the Sentiment Benchmarks with SentiWordNet

SWN Scores and Ratios of the Obesity Datasets

For the Obesity subsets, we calculated the overall sentiment by taking the difference between the positive score and negative scores of each subset ($Sc_{overall(T)} = Sc_{neg(T)} - Sc_{pos(T)}$). Therefore, subsets with a smaller overall score contain a more positive sentiment while subsets with a larger overall score contain a more negative sentiment. Among the obesity subsets, subset 6 (hypertension not obesity) contains the most positive overall sentiment (0.018) and subset 4 (obesity and hypertension) contains the most negative overall sentiment (0.026). The results of SWN analysis, reported in Table 4 (sorted by ascending overall score), show that majority of terms for each subset are objective (0.877).

Table 4 SWN sentiment scores of the obesity datasets

ID	Positive	Negative	Objective	Overall (negative–positive)
6	0.052	0.07	0.878	0.018
9	0.051	0.071	0.879	0.02
12	0.052	0.073	0.876	0.021
3	0.051	0.072	0.877	0.021
10	0.05	0.071	0.879	0.021
2	0.051	0.072	0.878	0.021
5	0.051	0.073	0.876	0.022
7	0.051	0.072	0.877	0.022
1	0.048	0.073	0.879	0.022
8	0.049	0.072	0.879	0.023
11	0.05	0.074	0.876	0.023
4	0.049	0.075	0.877	0.026
Average	0.05	0.072	0.877	0.022

We performed Welch's *t* test on the overall scores of each subset to evaluate whether the results were statistically significant [13]. Since our subsets are of different sizes and we cannot assume equal variance between subsets, we performed an independent two-sample Welch's *t* test on the overall sentiment score for each pair of subsets, with a null hypothesis stating that the means of both samples are equal ($\mu 1 = \mu 2$). The Welch *t* test shows that the difference in score between the most negative subset 4, obesity and hypertension, and the most positive subset 6, hypertension and not obesity, has a *p* value of 0.061, indicating a 93.9% chance of being statistically significant. Thus, we can confidently reject the null hypothesis and state that the difference between these subsets is significant.

The Welch's *t* test results also show that the largest *p* value of 0.989 occurs between subsets 3, diabetes, and subset 12, diabetes and not hypertension, having only a 1.1% chance of being statistically significant. This result aligns with the overall SWN sentiments scores, as both these subsets have an overall score of 0.021. Therefore, for these two subsets, we accept the null hypothesis and agree that the means of both subsets are equal.

The SWN sentiment ratios, computed for the subsets and reported in Table 5 (sorted by ascending percentage of negative terms), show that the results are stable across the sets but with small differences in the percentage of positive and negative terms. For example, subset 5 (obesity not hypertension) contains the highest percentage of positive terms (49.3%) along with the lowest percent of negative terms (50.7%), thus making it the most positive subset. In some cases, scores and ratios results support each other.

[3] https://www.nltk.org/.

[4] https://www.nltk.org/book/ch02.html.

[5] http://www.daviddlewis.com/resources/testcollections/reuters21578/.

Table 5 SWN sentiment ratios of the obesity datasets

ID	Num positive	Num negative	% positive	% negative
5	2636	2708	0.493	0.507
6	4322	4447	0.493	0.507
12	2931	3016	0.493	0.507
9	4321	4467	0.492	0.508
3	4941	5114	0.491	0.509
10	4597	4768	0.491	0.509
8	3006	3122	0.491	0.509
2	4971	5163	0.491	0.509
7	3503	3657	0.489	0.511
11	3186	3338	0.488	0.512
1	4142	4341	0.488	0.512
4	3744	3948	0.487	0.513
Average	3858.33	4007.42	0.491	0.509

Table 6 SWN sentiment scores for the Reuters-Science data

Dataset	Positive	Negative	Objective	Overall (negative–positive)
Crypto	0.060	0.049	0.890	−0.011
Space	0.049	0.039	0.912	−0.010
Electronics	0.056	0.046	0.898	−0.010
Med	0.066	0.061	0.872	−0.005
Average	0.058	0.049	0.893	−0.009

Table 7 SWN sentiment ratios for the Reuters-Science datasets

Datasets	Num positive	Num negative	% positive	% negative
Med	71,019	69,175	50.7	49.3
Electronics	45,278	43,947	50.7	49.3
Space	74,604	71,826	50.9	49.1
Crypto	78,341	75,301	51.0	49.0
Average	67,311	65,062	50.8	49.2

For example, subset 6 (hypertension not obesity) is the second most positive subset according to SWN ratios (49.3% positive, 50.7% negative), it is also the most positive subset according to SWN scores. Subset 4 (obesity and hypertension) is the most negative subset according to SWN ratios (48.7% positive, 51.3% negative) and the most negative subset according to SWN scores.

SNW Scores and Ratios of the Reuters-Science Datasets

We repeat the same experiments for the Science texts. The score results, reported in Table 6, show that majority of the terms for each dataset are objective (89.24%), which is very similar to what we observed in the Obesity dataset (87.74%). Observing the positive and negative scores for each subset, we see that the subset *Crypto* is the most positive (score = −0.011), and the subset *Med* is the most negative (score = −0.005).

The calculated SWN sentiment ratios are almost equal (Table 7) among the subsets, with approx. 1% difference between positive and negative terms. This small difference is akin to the difference exhibited by the Obesity dataset results. For example, the subset *Crypto* contains the highest percent of positive terms (51.0%) along with the lowest percent of negative terms (49.0%), making it the most positive subset. The subset *Med* contains the lowest percent of positive terms (50.7%) and the highest percent of negative terms (49.3%), making it the most negative subset.

We again performed an independent two-sample Welch's t test on the overall sentiment score for each pair of subsets, with a null hypothesis stating that the means of both samples are equal ($\mu 1 = \mu 2$). The Welch's t test results on the Reuters-Science datasets do not indicate a significant difference between the results, hence, not reported here.

Sentiment Analysis with Word2Vec and Doc2Vec

The Obesity Dataset

We trained a Word2Vec model on all the Obesity subsets, and then computed the cosine distance between each dataset. For the Doc2Vec evaluation, we also trained a Doc2Vec model on all the documents in the dataset. We then created an inferred vector for each of the subsets, and then calculated the cosine similarities between each of the inferred vectors. The Word2Vec similarities, reported by Table 8, indicate that similarity of the terms in each subset to all the other subsets is high. The subsets with the highest cosine similarity are subset 10 and subset 3; their cosine similarity is 99.97. This result is supported by both the SWN scores and ratios, as these both have the same overall sentiment score of 0.021 and the same SWN ratios (49.1% positive, 50.9% negative). The subsets with the lowest cosine similarities are subset 5 and subset 4, with a 97.22 cosine similarity. This result is supported by the SWN ratios, as subset 5 contains the highest percentage of positive terms (49.3%) and the lowest percent of negative terms (50.7%), which subset 4 contains the lowest percentage of positive terms (48.7%) and the highest percent of negative terms (51.3%).

The results of the Doc2Vec similarities (Table 9) show that the highest cosine similarity occurs again between subset 10 and subset 3 (0.9623). This result is supported by both the SWN scores and the Word2Vec cosine similarity results.

Table 8 Word2Vec cosine similarity scores (×100) on the obesity subsets

ID	1	2	3	4	5	6	7	8	9	10	11	12
1	100.0	99.87	99.77	99.86	98.31	99.55	99.89	99.66	99.54	99.80	99.47	98.92
2	99.87	100	99.94	99.72	98.20	99.87	99.81	99.45	99.86	99.96	99.53	99.16
3	99.77	99.94	100.0	99.57	98.28	99.87	99.82	99.13	99.95	99.97	99.22	99.42
4	99.86	99.72	99.57	100.0	97.22	99.22	99.85	99.36	99.27	99.72	99.12	98.21
5	98.31	98.20	98.28	97.22	100.0	98.55	97.88	98.55	98.36	97.91	98.55	99.25
6	99.55	99.87	99.87	99.22	98.55	100.0	99.46	99.19	99.95	99.80	99.48	99.49
7	99.89	99.81	99.82	99.85	97.88	99.46	100.0	99.17	99.58	99.87	98.99	98.93
8	99.66	99.45	99.13	99.36	98.55	99.19	99.17	100.0	98.96	99.15	99.82	98.38
9	99.54	99.86	99.95	99.27	98.36	99.95	99.58	98.96	100.0	99.88	99.20	99.55
10	99.80	99.96	99.97	99.72	97.91	99.80	99.87	99.15	99.88	100.0	99.20	99.13
11	99.47	99.53	99.22	99.12	98.55	99.48	98.99	99.82	99.20	99.20	100.0	98.62
12	98.92	99.16	99.42	98.21	99.25	99.49	98.93	98.38	99.55	99.13	98.62	100.0

Table 9 Doc2Vec cosine similarity scores (×100) on the obesity subsets

ID	1	2	3	4	5	6	7	8	9	10	11	12
1	100.0	79.43	81.16	92.29	68.42	56.38	89.76	79.73	66.70	77.50	66.81	55.48
2	79.43	100.0	89.21	76.97	73.49	79.27	77.81	75.38	81.22	89.50	63.08	66.43
3	81.16	89.21	100.0	82.80	62.41	72.33	82.48	62.92	78.66	96.23	51.21	68.06
4	92.29	76.97	82.80	100.0	61.46	58.68	94.30	78.33	70.31	80.26	60.45	53.41
5	68.42	73.49	62.41	61.46	100.0	61.80	67.74	65.85	66.41	62.13	70.58	72.88
6	56.38	79.27	72.33	58.68	61.80	100.0	63.35	57.87	90.20	72.81	65.02	64.70
7	89.76	77.81	82.48	94.30	67.74	63.35	100.0	74.84	71.91	79.04	66.60	60.74
8	79.73	75.38	62.92	78.33	65.85	57.87	74.84	100.0	60.62	67.24	77.22	35.46
9	66.70	81.22	78.66	70.31	66.41	90.20	71.91	60.62	100.0	78.42	61.92	7292
10	77.50	89.50	96.23	80.26	62.13	72.81	79.04	67.24	78.42	100.0	55.20	67.08
11	66.81	63.08	51.21	60.45	70.58	65.02	66.60	77.22	61.92	55.20	100.0	48.14
12	55.48	66.43	68.06	53.41	72.88	64.70	60.74	35.46	72.92	67.08	48.14	100.0

Table 10 Reuters-Science Word2Vec cosine similarity scores

	Crypto	Electronics	Med	Space
Crypto	1.000	0.690	0.611	0.542
Electronics	0.690	1.000	0.727	0.606
Med	0.611	0.727	1.000	0.638
Space	0.542	0.606	0.638	1.000

Table 11 Reuters-Science Doc2Vec cosine similarity scores

	Crypto	Electronics	Med	Space
Crypto	1.000	0.128	−0.074	−0.320
Electronics	0.128	1.000	0.379	0.070
Med	−0.074	0.379	1.000	0.325
Space	−0.320	0.070	0.325	1.000

The lowest cosine similarity occurs between subset 12 and subset 8 (35.46). This result is not strongly supported by SWN scores or ratios.

The Reuters Dataset

For the Word2Vec cosine similarity on the Reuters dataset, shown in Table 10, we see the subsets *Med* and *Electronics* have the highest cosine similarity (0.727). However, both SWN scores and ratios do not support this similarity. The lowest cosine similarity score occurs between *Space* and *Crypto* (0.542), which is also not supported by SWN scores or ratios.

The results of the Doc2Vec cosine similarities (Table 11) show that the highest cosine similarity occurs between the subsets *Med* and *Electronics* (0.379), but this is not supported by the results in SWN. The lowest cosine similarity occurs between the subsets *Crypto* and *Space* (−0.32), which is also not supported by the results of SWN. Note that the Word2Vec model showed the same result for both the highest and lowest cosine similarities. Interestingly, the third lowest cosine similarity observed is between *Crypto*

Table 12 Correspondence between Word2Vec, Doc2Vec, and SWN results

Data sets	Word2Vec		Doc2Vec	
	Subsets with SWN sentiment highest similarity	Subsets with SWN sentiment lowest similarity	Subsets with SWN highest sentiment similarity	Subsets with SWN lowest sentiment similarity
Obesity[a]	✓	✓	✓	×
Reuters	×	×	×	×

✓ supported by SWN results, × not supported by SWN results

[a]Indicates the data set with significant difference between the most positive and most negative SWN evaluations.

and *Med* (−0.074), which are the two most different subsets in sentiment according to SWN.

Correspondence with the SWN results

We note that the Welch test results serve as a strong predictor for correspondence of the Word2Vec sentiment evaluation to the SWN evaluation:

- For the Obesity subsets, the Welch *t* test showed significant difference between the SWN most positive and most negative sentiment evaluations. In this case, the Word2Vec and the SWN sentiment evaluation correspond to each other.
- For the Reuters data sets, the Welch *t* test did not find difference between the SWN most positive and most negative results to be significant. For this data set, the Word2Vec results do not correspond to the SWN results.

In Table 12, we summarize correspondence between Word2Vec and Doc2Vec results and those obtained by SWN.

The Study Limitations and Future Work

Our empirical results are obtained on datasets of medical and scientific text, split into several subsets. Legal texts, another type of professional text [30], would be a natural extension of our study; however, they remain out of its current scope. To extend unsupervised sentiment analysis to legal documents, we need to build in-depth linguistic models to process legal terminology embedded into professional sublanguage and factual description of events; complex sentences structures bring in additional challenges as well as archaic, formal, formulaic, unusual words and word forms, overuse of the conjunctions, impersonal constructions, and overuse of passive voices.

The current sentiment benchmark values have been obtained using one, albeit advanced, lexical resource (i.e., SWN). We plan to include additional lexical resources (e.g., MPQA [15]) as a part of our analysis. Another possible extension of the current work is to compare the impact of semantic similarity measures on the sentiment analysis; review of the semantic similarity measures can be found in [1].

We were not able to clearly observe positive results from the Doc2Vec model, where performance relies on complete documents rather than individual words. We hypothesise that (a) for a stronger generalization of the Doc2Vec performance, researchers want to apply the tool on longer and more informative documents than supplied by the current data sets. Short and succinctly written texts may not be suitable for building reliable document embeddings; (b) to obtain more conclusive results, Doc2Vec needs a richer and more diversified vocabulary than the current texts had provided; in other words, the types to tokens ratio should be higher than in the current data sets (Table 1 reports the corresponding results). (c) to circumvent the above-mentioned deficiencies, Doc2Vec can be used in a semi-supervised setting, i.e., it can be pre-trained on a small set of labelled examples and then applied on a bigger set of unlabelled examples. This would result in more accurate training on the context of positive/negative sentiments, as well as a more reliable performance for Doc2Vec.

Conclusions

In this study, we have proposed a novel approach to unsupervised sentiment analysis of medical and scientific texts. We have obtained the empirical results on medical (i2b2) and scientific (Reuters-Science) documents. The texts did not have sentiment labels or any other sentiment-related annotations. We have chosen the obesity data since obesity is a highly stigmatized condition in society, with obese individuals often being discriminated against by society [29]. It would be important to observe if physicians and nurses exhibit the same bias in their clinical writings. In the medical and scientific texts, sentiments are stated indirectly and subtly expressed, thus making manual annotation unreliable and supervised learning unfeasible. We have shown that sentiment analysis in such cases can be done using appropriate lexical resources as the source of the benchmark evaluation (SentiWordNet) and then employing statistical assessment (Welch's test) of the results obtained by the analytics algorithms.

To summarize: utilizing state-of the art techniques, we applied unsupervised machine learning models Word2Vec and Doc2Vec to detect the expressed sentiments in the medical and scientific texts. In unsupervised sentiment analysis of the Obesity data, the Word2Vec results correspond to those of SentiWordNet. The obtained Doc2Vec results in part

correspond to the Word2Vec and the SentiWordNet results. On the same data, the Welch's test indicated significant difference between the most positive and most negative SWN sentiment evaluation. On the Reuters data, the Welch test did not find significant difference between the most positive and most negative SWN evaluation, and the Word2Vec and Doc2Vec results do not correspond to those of SWN.

We conclude that in unsupervised sentiment analysis of medical texts, Word2Vec exhibited a more reliable sentiment analysis than Doc2Vec. The Welch's test significance results can serve as an indicator of correspondence between Word2Vec and SWN results. A more detailed study into the performance of Doc2Vec is needed.

Acknowledgements The authors thank participants of the 32rd Canadian Conference on AI (Canadian AI 2019) for a helpful discussion of the study. The authors thank the anonymous SN Computer Science reviewers for useful comments on the 1st submission of this article.

Author contributions QC and MS equally contributed to the study design and ideas implemented in this project. They both wrote parts of the text. QC ran the experiments. MS supervised the study.

Funding A preliminary part of this study was supported by the Discovery program of NSERC Canada.

Availability of data and material The data sets are available from http://www.daviddlewis.com/resources/testcollections/reuters21578/ and https://www.i2b2.org/NLP/DataSets/Main.php.

Code availability We have adapted open source software Gensim: Topic modelling for humans (radimrehurek.com).

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Araque O, Zhu G, Iglesias CA. A semantic similarity-based perspective of affect lexicons for sentiment analysis. Knowl Based Syst. 2019;165:346–59.
2. Aslam F, Awan T, Syed JH, Kashif A, Parveen M. Sentiments and emotions evoked by news headlines of coronavirus disease (COVID-19) outbreak. Humanit Soc Sci Commun. 2020;7(1):1–9.
3. Bobicev V, Sokolova M. Inter-annotator agreement in sentiment analysis: machine learning perspective. In: RANLP 2017. ACL. 2017. p. 97–102.
4. Carducci G, Leontino M, Radicioni DP, Bonino G, Pasini E, Tripodi P (2019) Semantically aware text categorisation for metadata annotation. In: Italian research conference on digital libraries. Springer. p. 315–30.
5. Chen Q, Sokolova M.Word2vec and doc2vec in unsupervised sentiment analysis of clinical discharge summaries. 2018. arXiv:1805.00352.
6. Chen Q, Sokolova M. Unsupervised sentiment analysis of objective texts. In: Canadian conference on artificial intelligence. Springer. 2019. p. 460–65.
7. Das S, Mandal SK, Basu A. Mining multiple informational text structure from text data. In: ICCIDS 2019. Procedia Computer Science. 2020. p. 2211–20.
8. Deep Learning for Java. Word2Vec, Doc2vec & GloVe: Neural Word Embeddings for Natural Language Processing. Deep Learning for Java. 2017. https://deeplearning4j.org/word2vec.html.
9. Delacre M, Lakens D, Leys C. Why psychologists should by default use Welch's t-test instead of Student's t-test. Int Rev Soc Psychol. 2017;30(1). https://www.ripsirsp.com/articles/10.5334/irsp.82/
10. Denecke K, Deng Y. Sentiment analysis in medical settings. Artif Intell Med. 2015;64(1):17–27.
11. Deng Y, Declerck T, Lendvai P, Denecke K. The generation of a corpus for clinical sentiment analysis. In: The semantic web—ESWC 2016 satellite events. 9989. Cham: Springer; 2016.
12. Deng Y, Stoehr M, Denecke K. Retrieving attitudes: sentiment analysis from clinical narratives. In: Medical information retrieval workshop at SIGIR 2014. 2014. p. 12–5.
13. Derrick B, Toher D, White P. Why Welchs test is Type I error robust. Quant Methods Psychol. 2016;12(1):30–8.
14. Esuli A, Sebastiani F. SENTIWORDNET: a publicly available lexical resource for opinion mining. In: LREC'06. 2006. p. 417–22.
15. Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision. Stanford. 2009.
16. Herrmannova D, Young S, Patton R, Stahl C, Kleinstreuer N, Wolfe M. Unsupervised identification of study descriptors in toxicology research: an experimental study. In: International workshop on health text mining and information analysis. ACL. 2018. p. 71–82.
17. Jin X, Xu Y. Research on the sentiment analysis based on machine learning and feature extraction algorithm. In: 2019 IEEE 10th international conference on software engineering and service science (ICSESS). IEEE. 2019. p. 366–69.
18. Lakens D. Always use Welch's t-test instead of Student's t-test. The 20% Statistician. 2015. http://daniellakens.blogspot.ca/2015/01/always-use-welchs-t-test-instead-of.html. Accessed 23 Apr 2018.
19. Le Q, Mikolov T. Distributed representations of sentences and documents. ICML. 2014;32:1188–96.
20. Lennox RJ, Veríssimo D, Twardek WM, Davis CR, Jarić I. Sentiment analysis as a measure of conservation culture in scientific literature. Conserv Biol. 2020;34(2):462–71.
21. Losada DE, Gamallo P. Evaluating and improving lexical resources for detecting signs of depression in text. Lang Resour Eval. 2020;54(1):1–24.
22. Liu S, Lee I. Extracting features with medical sentiment lexicon and position encoding for drug reviews. Health Inf Sci Syst. 2019;7(1):11.
23. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. 2013. arXiv:1301.3781 [CoRR/abs].
24. Mishra S, Pappu A, Bhamidipati N. Inferring advertiser sentiment in online articles using wikipedia footnotes. In: The 2019 World Wide Web conference. 2019. p. 1224–31.
25. Mohan S, Guha A, Harris M, Popowich F, Schuster A, Priebe C. The impact of toxic language on the health of Reddit communities. In: Canadian conference on artificial intelligence. Springer; 2017. p. 51–6.
26. Majumder N, Hazarika D, Gelbukh A, Cambria E, Poria S. Multimodal sentiment analysis using hierarchical fusion with context modeling. Knowl Based Syst. 2018;161:124–33.
27. Naseem U, Razzak I, Musial K, Imran M. Transformer based deep intelligent contextual embedding for twitter sentiment analysis. Futur Gener Comput Syst. 2020;113:58–69.

28. Nakov P, Kozareva Z, Ritter A, Rosenthal S, Stoyanov V, Wilson T. Semeval-2013 task 2: sentiment analysis in Twitter. In: Joint conference on lexical and computational semantics, vol 2. Association for Computational Linguistics; 2013. p. 312–20.
29. Puhl R, Heuer C. The stigma of obesity: a review and update. Obesity. 2009;17(5):941–64.
30. Queudot M, Meurs MJ. Artificial intelligence and predictive justice: limitations and perspectives. In: International conference on industrial, engineering and other applications of applied intelligent systems. Cham: Springer; 2018. p. 889–97.
31. Řehůřek R, Sojka P. Software framework for topic modelling with Large Corpora. In: The LREC workshop on new challenges for NLP frameworks. 2010.
32. Řehůřek R, Sojka P. Gensim—statistical semantics in python. 2011.
33. Sokolova M, Bobicev V. What sentiments can be found in medical forums? In: Proceedings of the international conference recent advances in natural language processing RANLP 2013. 2013. p. 633–39.
34. Sokolova M. Big text advantages and challenges: classification perspective. Int J Data Sci Anal. 2018;5(1):1–10.
35. Taboada M. Sentiment analysis: An overview from linguistics. Annu Rev Linguist. 2016;2:325–47.
36. Tafti AP, Wang Y, Shen F, Sagheb E, Kingsbury P, Liu H. Integrating word embedding neural networks with PubMed abstracts to extract keyword proximity of chronic diseases. In: IEEE EMBS. 2019.
37. Taghizadeh N, Doostmohammadi E, Seifossadat E, Rabiee HR, Tahaei MS SINA-BERT: a pre-trained language model for analysis of medical texts in Persian. 2021. arXiv:2104.07613.
38. Tang D, Wei F, Yang N, Zhou M, Liu T, Qin B. Learning sentiment-specific word embedding for twitter sentiment classification. In: ACL. ACL. 2014.
39. Trusca M. Efficiency of SVM classifier with Word2Vec and Doc2Vec models. In: International conference on applied statistics. 2019. p. 496–503.
40. Uzuner Ö. Recognizing obesity and co-morbidities in sparse data. JAMIA. 2009;16(4):561–70.
41. van Zoonen W, van der Toni GL. Social media research: the application of supervised machine learning in organizational communication research. In: Computers in human behavior, 2016. p. 132–41.
42. Wagstaff K, Francis R, Gowda T, Lu Y, Riloff E, Singh K, Lanza N. Mars target encyclopedia: rock and soil composition extracted from the literature (No. LA-UR-18-21439). Los Alamos National Lab (LANL), USA. 2018.
43. Wang T, Lu K, Chow KP, Zhu Q. COVID-19 sensing: negative sentiment analysis on social media in China via BERT model. IEEE Access. 2020;8:138162–9.
44. Wang Y, Sun A, Han J, Liu Y, Zhu X. Sentiment analysis by capsules. In: World Wide Web conference. 2018. p. 1165–74.
45. Xie R, Chu SKW, Chiu DKW, Wang Y. Exploring public response to COVID-19 on Weibo with LDA topic modeling and sentiment analysis. Data Inf Manag. 2021;5(1):86–99.
46. Yang T, Yao R, Yin Q, Tian Q, Wu O. Mitigating sentimental bias via a polar attention mechanism. Int J Data Sci Anal. 2021;11(1):27–36.
47. Yousef R, Tiun S, Omar N, Alshari E. Enhance medical sentiment vectors through document embedding using recurrent neural network. In: IJACSA. 2020. p. 372–78.
48. Yousif A, Niu Z, Tarus JK, Ahmad A. A survey on sentiment analysis of scientific citations. Artif Intell Rev. 2019;52(3):1805–38.
49. Zeng Z, Zhou W, Liu X, Lin Z, Song Y, Kuo MD, Chiu WHK. A variational approach to unsupervised sentiment analysis. 2020. arXiv:2008.09394.