Evaluating Semantic Similarity for Adverse Drug Event Narratives

H. Irfan Khaja[1], Marie Abate [2], Wanhong Zheng [3], Ahmed Abbasi [4], Donald Adjeroh[1]

[1]Dept. of Comp. Sci. & Elec. Engg., West Virginia University, Morgantown, WV 26506, USA
[2]School of Pharmacy, West Virginia University, Morgantown, WV 26506, USA
[3]School of Medicine, West Virginia University, Morgantown, WV 26506, USA
[4]McIntire School of Commerce, University of Virginia, Charlottesville, VA 22904, USA
hkirfan@mix.wvu.edu, don@csee.wvu.edu

Abstract. We propose a method to evaluate adverse drug event (ADE) narratives using biomedical semantic similarity measures. Automated drug surveillance systems have used social media as a prime resource to detect ADEs. However, the problem of language usage over social media has been a challenge in evaluating the performance of such systems. We address this key issue by using semantic similarity measures and the biomedical vocabularies from the Unified Medical Language System. This is important in comparing results of social media driven approaches against standard reference documents from regulatory agencies.

Keywords: Semantic Similarity, Adverse Drug Events, Social Media.

1 Introduction

High morbidity and mortality rates are associated with adverse drug events (ADEs), and hence, pharmacovigilance serves a critical task in post marketing surveillance. Recent advancements have shown a good potential for the detection of ADEs using social media much earlier than the traditional reporting systems [1]–[6]. Unfortunately, most of the work on detecting ADEs through social media have not emphasized the issue of language usage. The language used in expressing issues by healthcare consumers on social media forums and microblogging websites like Twitter is often very casual and informal [7]. On the other hand, warning labels and notifications from official regulatory agencies (such as the Food and Drug Administration (FDA) in the US) are formal documents and usually described in a language that is very carefully selected by biomedical experts. This raises a major concern as the words detected from social media channels by the surveillance systems do not exactly match with the contents of a typical FDA Black Box Warning (BBW) label or alert notification.

For many pairs of terms, there is a potential to miss the semantic similarity between social media extracted ADE terms and terms from FDA notification when two sets of terms do not share exact text. More specifically the problem is as follows: given a formal FDA ADE narrative: X= $\{x_1, x_2, \dots x_n\}$, and an informal ADE narrative from social media Y= $\{y_1, y_2, \dots y_m\}$, determine the semantic similarity between X and Y. The three major issues related to semantic similarity in automated drug surveillance are: 1) How to measure semantic similarity between social media narratives and official formal documents, 2) How to use semantic similarity to evaluate the accuracy of detected

ADEs, and 3) How to use semantic similarity to improve ADE signal detection. This work focuses on the first two problems. In general, X and Y could represent any two documents with words. Thus, semantic similarity can also have applications in other fields like medical appliances, ecommerce, etc.

Previously, Yang et al. [4] attempted to address the problem of health consumers' language over the Internet by generating ADE lexicons using Consumer Health Vocabulary (CHV) [7]. But, this did not address the issue comprehensively, as there are over 200 biomedical vocabularies in just UMLS (Unified Medical Language System), which also includes CHV [8]. Here, we use UMLS-Similarity program developed by McInnes et al. [9], for computing semantic similarity. It incorporates well-known semantic similarity and semantic relatedness measures. The prominent ones include path finding measures (such as Rada et al. [10], and Wu & Palmer [11]) as well as information content (IC) measures (such as Jiang & Conrath [12], and Sánchez et al. [13]). In prior work, Park et al. evaluated vocabularies from UMLS based on diabetes-related terms extracted from social media [14]. However, it confines itself to only one subset of the vast healthcare domain. We aimed at evaluating all the measures listed in UMLS-Similarity and vocabularies in UMLS to determine the best combination of measures and vocabulary in computing semantic similarity for ADE narratives.

2 Materials and Methods

Our methodology follows the procedure: 1) Identify the best vocabulary configurations (VCs) to use, 2) Determine the best combination of VCs and similarity measurement algorithms (SMAs) via joint optimization, and 3) Perform semantic similarity measurement using VC and SMA on given narratives.

2.1 Datasets

Problem Domain Terms. To evaluate VCs and SMAs, we used a list of terms grouped into anatomy and reaction categories. This dataset was earlier used by Adjeroh et al. [2] to study ADEs using social media data. The dataset has 105 anatomy terms and 202 reaction terms (called clusters). Each cluster was expanded with words having similar meanings, resulting in a new list with 178 anatomy terms, and 417 reaction terms.

Human Ratings. Language is a major concern in evaluating the signals generated from social media, hence, the testing on SMAs and VCs should be based on the ratings obtained from general healthcare consumers along with healthcare professionals. Thus, we used human ratings as the standard to compare the performance of each combination of SMA and VC. Initially, we had 178 anatomy terms and 417 reaction terms, and forming pairs with all these terms would lead to over 100,000 pairs and that would have been impossible for the respondents to rate the similarity. Thus, we randomly selected 30 anatomy terms forming a set of 435 [(30*29)/2] anatomy pairs and 40 reaction pairs forming a set of 780 [(40*39)/2] reaction pairs. Further, to rate these 1215 pairs we contacted 6 computer science graduate researchers having appreciable knowledge of biomedical vocabulary usage over social media. Finally, based on their ratings a template with a set of 100 pairs was designed comprising 50 anatomy pairs and 50 reaction pairs. This template had rating options 0, 0.25, 0.5, 0.75 and 1 indicating levels from non-similar to very similar. We obtained 130 user ratings across the United States. This consists of 54 individuals coming from 5 different universities with health sciences and engineering background, and 76 from Amazon Mechanical Turk users having at least US Bachelor's degree. Further, we selected 117 ratings by excluding the outliers that had a negative correlation with the mean. We also analyzed the inter-rater agreement in terms of average correlation between raters. We filtered the ratings to achieve the benchmark of 80% average correlation and this resulted in a total of 107 ratings.

FDA BBW. To evaluate our work, we used FDA black box warning (BBW) labels as gold standard references and extracted ADE terms from the labels from January 2008 to April 2015. (http://www.fda.gov/safety/medwatch/safetyinformation/). This included 107 BBWs, on 90 drugs over the seven-year period.

2.2 Selection of Vocabulary Configurations (VCs)

Since the biomedical terms are found in multiple vocabularies it becomes a challenging question to decide which vocabulary to be used. The harder part is to find how good a given vocabulary is, in terms of covering all terms in a given problem domain.

Initial Selection. UMLS has a huge collection of over 200 biomedical vocabularies which serves as a good resource for our work. However, we cannot use all the vocabularies in UMLS-Similarity due to performance and computational issues (see [15] for example). For our domain-specific social media extracted ADE terms, we followed the discussions in Park et al. [14], and selected vocabularies represented by source abbreviation (SAB): SNOMEDCT_US, CHV, MSH, LCH_NW, LNC, RXNORM, NCI_FDA, VANDF, and MTHSPL from UMLS [8]. The work in [14] was based on terms extracted from social media using queries for terms related to diabetes. For a more comprehensive treatment, we have considered some additional vocabularies where the content is closely related to ADE terms; namely, FMA, MDR, UWDA, WHO, NCI_NICHD, NCI_CTCAE, NDFRT_FDASPL, ICD10CM, MTHHH, and GS. Thus, given our specific problem domain of analyzing ADEs over social media channels, we had a total of 19 vocabularies to start our study.

Refining the VCs selection. Our next task is to reduce the list to get the best possible VCs based on the concepts. We considered the following features:

1. Total CUI's: Total # of concept unique identifiers (CUIs) listed for the vocabulary;
2. Terms detected: number of problem domain terms detected in the vocabulary;
3. Concept coverage: number of concepts (CUI's) listed for problem domain terms;
4. Unique concepts: number of unique CUIs listed for each vocabulary; and
5. Clusters detected: number of clusters which had at least one term detected as CUI.

For our purpose, a good vocabulary is expected to have higher values for these features.

2.3 Similarity Measurement Algorithms (SMAs)

For automated evaluation of semantic similarity, vocabulary is just one piece of the puzzle. Another key piece is the specific SMA to be used to measure the similarity using the identified VC. Thus, having narrowed down the VCs we now turn to the problem of selecting the SMAs. Interestingly, the match performance can also be influenced by the vocabulary used. Thus, the final choice of vocabulary cannot be made in isolation but must consider the specific SMA being used. We used all SMAs in UMLS-Similarity except the *vector* measure which is meant to compute relatedness (see Table 1).

Joint Selection of VC and SMA. We computed similarity values for the problem domain terms using each combination of selected VCs and the SMAs. To select the best SMA and VC, we compared their results with those of human observers in two steps: 1) using Pearson correlation against the mean rating from human observers, and 2) using information retrieval measures, where we grouped the problem domain term pairs into 3 classes: **similar pairs**, **unknown pairs**, and **non-similar pairs**. Let $S(x, y)$ be the semantic similarity value between term pair (x, y), as returned by a given algorithm. We then used two thresholds τ_1 and τ_2 ($\tau_1 \geq \tau_2$) to classify a word pair (v_1, v_2):

$$Class\big(S(v_1,v_2)\big) = \begin{cases} similar, & S(v_1,v_2) > \tau_1 \\ unknown, & \tau_1 \geq S(v_1,v_2) \geq \tau_2 \\ not\ similar, & S(v_1,v_2) < \tau_2 \end{cases} \quad (1)$$

Table 1. Similarity measurement algorithms in UMLS-Similarity. *References for each can be found in [9].

#	UMLS-Similarity Notation	Type	#	UMLS-Similarity Notation	Type
1	***lch***	path finding	9	***lin***	IC based
2	***wup***	path finding	10	***jcn***	IC based
3	***zhong***	path finding	11	***vector***	context vector
4	***path***	path finding	12	***pks***	path finding
5	***upath***	path finding	13	***faith***	IC based
6	***cdist***	path finding	14	***cmatch***	feature based
7	***nam***	path finding	15	***batet***	feature based
8	***res***	IC based	16	***sanchez***	IC based

We used traditional information retrieval measures, namely, Precision (Pr), Recall (Rc), and F-measure (Fm) to evaluate the performance of combinations of VCs and SMAs across the three classes.

3 Experiments and Results

3.1 Filtering Vocabularies

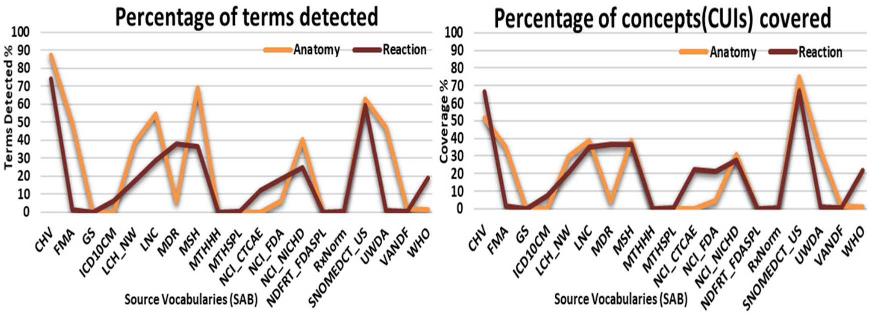
Using programs from the UMLS-Interface [9], we listed all the Concept Unique Identifiers (CUIs) for vocabularies configured with various relations defined in UMLS [8]. Interestingly, some vocabularies have concepts but are not connected by any relations. Additionally, we obtained the CUIs for all the problem domain terms to evaluate each vocabulary based on various features discussed in section 2.2. Fig. 1 shows some of the features used to describe the vocabularies. We observed that the top 5 vocabularies for anatomy category are SNOMEDCT, CHV, LNC, MSH, and FMA. The top 5 vocabularies for reaction category are SNOMEDCT, CHV, MDR, MSH, and LNC. However, we found that, CHV has no relations defined between CUIs which restricts its use independently. Thus, we used CHV in combination with other VCs as it has more coverage of terms, and has been shown to improve performance [4], [14].

3.2 Joint Selection of VC and SMA

Correlation Analysis. If the significance level is ≤ 5% (i.e., p-value ≤ 0.05) and the corresponding correlation coefficient is positively high for any VC and SMA, then we say that the SMA or VC is favored. From Table 2 and Fig. 2, we can see that for anatomy category the SMAs which frequently appear to be good are *cmatch, jcn* and *sanchez* with VCs CHV-SNOMEDCT and CHV-LNC. For reaction category, we did not get significant p-value to favor any of the algorithms. However, it has been observed that *nam* has very high correlation coefficient with vocabularies CHV-MDR and CHV-MSH and undefined value for CHV-LNC. This behavior is because of the similarity values being -1.0 for most term pairs, resulting in less variability. Overall, the correlation analysis suggests that CHV-SNOMEDCT and CHV-MDR are the best VCs for working on reaction category terms (see Fig. 2(b)).

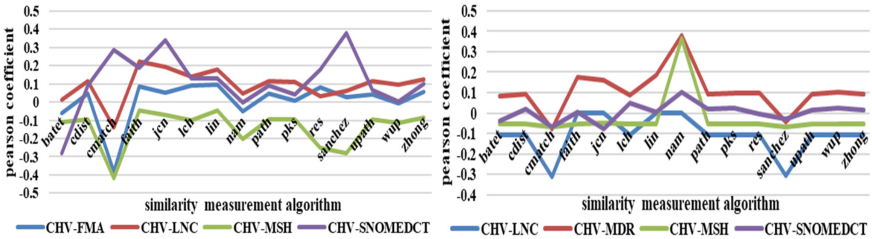
Table 2. Outcomes of Pearson correlation

SMA favored		VC favored	
Anatomy	**Reactions**	**Anatomy**	**Reactions**
cmatch, jcn, sanchez	*nam*	CHV-SNOMEDCT, CHV-LNC	CHV-SNOMEDCT, CHV-MDR



(a) Terms detected (b) Concepts Identified

Fig. 1. Features for filtering vocabularies based on problem domain terms.



(a) Anatomy (b) Reaction

Fig. 2. Correlation of computed similarity values with human rating

Information Retrieval Factors. For the median of human ratings, we chose thresholds τ_1 as 0.75 and τ_2 as 0.3 to classify them into similar pairs, unknown pairs, and non-similar pairs. Similar to human ratings, for the SMA-VC obtained similarity values we chose τ_1 ranging from 0.5 to 0.95 and τ_2 ranging from 0.05 to 0.45 with a step size of 0.05. We selected the top 5 SMA-VCs based on F-measure against human rating statistic.

Table 3. Top 5 SMA/VC (Similar pairs--Anatomy)

Measure	τ_1	τ_2	τ_{diff}	Configuration	Pr	Rc	Fm
jcn	0.8	0.5	0.3	CHV-SNOMEDCT	0.89	0.62	0.73
faith	0.7	0.5	0.2	CHV-SNOMEDCT	0.89	0.62	0.73
lin	0.8	0.45	0.35	CHV-SNOMEDCT	0.89	0.62	0.73
cmatch	0.5	0.45	0.05	CHV-SNOMEDCT	0.72	0.62	0.67
sanchez	0.8	0.5	0.3	CHV-SNOMEDCT	0.72	0.62	0.67

Table 4. Top 5 SMA/VC (Similar pairs--Reaction)

Measure	τ_1	τ_2	τ_{diff}	Configuration	Pr	Rc	Fm
pks	0.55	0.35	0.2	CHV-SNOMEDCT	1	0.3	0.46
res	0.8	0.3	0.5	CHV-MDR	1	0.3	0.46
sanchez	0.5	0.4	0.1	CHV-MDR	1	0.3	0.46
wup	0.75	0.3	0.45	CHV-SNOMEDCT	1	0.3	0.46
sanchez	0.85	0.3	0.55	CHV-SNOMEDCT	0.75	0.3	0.43

For anatomy terms (Table 3), we found that the SMAs *jcn, faith, lin, cmatch* and *sanchez* with CHV-SNOMEDCT VC are having high F-measure values with respect to human ratings. For reaction category (Table 4), the SMAs *wup, lin, pks, cmatch* with CHV-SNOMEDCT, and *res* with CHV-MDR VC performed well. Interestingly, we observe that *sanchez* has good F-measure for both CHV-SNOMEDCT and CHV-MDR.

3.3 Evaluating Narratives in ADE Surveillance Systems.

Considering both the information retrieval metrics and the correlation analysis, our results suggest the following: for anatomy term pairs, we should use *jcn, cmatch,* or *sanchez* SMA, with CHV-SNOMEDCT VC. For reaction term pairs, we should use *sanchez*, *wup*, or *res* SMA, with CHV-SNOMEDCT or CHV-MDR VC. A key observation is the need for a combination of vocabularies (typically, CHV with some others), rather than one single vocabulary as has been used in prior work, such as [4]. Prior work also did not consider the impact of the SMA on the results. We evaluated suggested ADE narratives from social media based on the method described in [16] using the BBW dataset (refer Section 2.1). We considered four cases: 1) exact match, i.e., not using semantic similarity; and the other 3 cases with SMA *sanchez* along with VC 2) CHV, 3) SNOMEDCT and 4) combination of CHV-SNOMEDCT (See Table 5). Clearly, our suggested approach using combination of CHV and SNOMEDCT performed better than others.

Table 5. Evaluating social media ADE narratives for BBW data

Approach	Anatomy			Reaction		
	Pr	Rc	Fm	Pr	Rc	Fm
exact match	0.048	0.176	0.076	0.022	0.140	0.038
CHV	0.048	0.176	0.076	0.024	0.141	0.041
SNOMEDCT	0.181	0.395	0.249	0.155	0.402	0.224
CHV-SNOMEDCT	**0.197**	**0.452**	**0.275**	**0.175**	**0.465**	**0.255**

4 Discussion and Conclusion

In this work, we chose UMLS-Similarity as it is built on UMLS which provides access to multiple vocabularies. The human ratings we used had a good representation of doctors, health professionals, health science students, engineering graduates and general graduate students. As the participants were familiar with social media as a significant source of healthcare information, we believe our dataset best fits the testing. We followed a-step-by-step approach evaluating all vocabularies and measures exhaustively, to get the best suitable VC-SMA combination for the ADE terms. Our results showed that, CHV-SNOMEDCT is the best VC for anatomy terms using the intrinsic IC-based measures *sanchez* or *jcn*. It is also observed that CHV-MDR and CHV-SNOMEDCT VCs work well for reaction category terms with *sanchez*. However, our results also indicate that using biomedical ontologies and the similarity measures is not sufficient for reaction category terms. The major reason is that reaction terms are more general and are not as specific when compared to anatomy category terms. Thus, we believe that using general English vocabularies such as WordNet [17] along with UMLS would improve the semantic similarity for reaction category terms. We plan to evaluate this in further studies. Our findings also show that the vocabulary MedDRA--Medical Dictionary for Regulatory Activities (abbreviated as MDR in UMLS) has a good representation of reaction category problem domain terms. This can be considered in the light of the fact that SIDER, a well-known dataset for representing side effects uses MedDRA to generate side effect names [18].

References

1. A. Sarker, R. Ginn, A. Nikfarjam, K. O'Connor, et al., "Utilizing social media data for pharmacovigilance: A review," *J. Biomed. Inform.*, vol. 54, pp. 202–212, 2015.
2. D. Adjeroh, R. Beal, A. Abbasi, W. Zheng, et al., "Signal Fusion for Social Media Analysis of Adverse Drug Events," *IEEE Intell. Syst.*, vol. 29, no. 2, pp. 74–80, 2014.
3. A. Abbasi, D. Adjeroh, M. Dredze, M. J. Paul, et al., "Social media analytics for smart health," *IEEE Intell. Syst.*, vol. 29, no. 2, pp. 60–80, 2014.
4. C. C. Yang, H. Yang, and L. Jiang, "Postmarketing Drug Safety Surveillance Using Publicly Available Health-Consumer-Contributed Content in Social Media," *ACM Trans. Manag. Inf. Syst.*, vol. 5, no. 1, pp. 1–21, 2014.
5. R. B. Correia, L. Li, and L. M. Rocha, "Monitoring potential drug interactions and reactions via network analysis of Instagram user timelines.," *Biocomput. Proc. Pacific Symp.,* vol. 21, pp. 492–503, 2016.
6. A. Abbasi, J. Li, S. Abbasi, D. Adjeroh, et al., "Don't Mention It? Analyzing User-Generated Content Signals for Early Adverse Drug Event Warnings," *Proceedings, Wkshp. Inf. Technol. Syst. (WITS), Dallas, TX.*, pp. 1–16, 2015.
7. Q. T. Zeng and T. Tse, "Exploring and developing consumer health vocabularies," *Journal of the American Medical Informatics Association*, vol. 13, no. 1. pp. 24–29, 2006.
8. National Library of Medicine (US), *UMLS® Reference Manual*. National Library of Medicine (US), 2009.
9. B. T. McInnes, T. Pedersen, and S. V. S. Pakhomov, "UMLS-Interface and UMLS-Similarity : open source software for measuring paths and semantic similarity.," *AMIA Annu. Symp. Proc.*, pp. 431–5, 2009.
10. R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," *IEEE Trans. Syst. Man Cybern.*, vol. 19, no. 1, pp. 17–30, 1989.
11. Z. Wu and M. Palmer, "Verbs semantics and lexical selection.," *Proc. 32nd Annu. Meet. Assoc. Comput. Linguist.* -, pp. 133–138, 1994.
12. J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," *Proc. Int. Conf. Res. Comput. Linguist. Taiwan*, 1997.
13. D. Sánchez, M. Batet, D. Isern, and A. Valls, "Ontology-based semantic similarity: A new feature-based approach," *Expert Syst. Appl.*, vol. 39, no. 9, pp. 7718–7728, 2012.
14. M. S. Park, Z. He, Z. Chen, S. Oh, and J. Bian, "Consumers' Use of UMLS Concepts on Social Media: Diabetes-Related Textual Data Analysis in Blog and Social Q&A Sites.," *JMIR Med. Informatics*, vol. 4, no. 4, p. e41, Nov. 2016.
15. "Re: [umls-similarity] Practical large coverage configuration." [Online]. Available: https://www.mail-archive.com/umls-similarity@yahoogroups.com/msg00334.html. [Accessed: 15-Mar-2018].
16. H. I. Khaja, "Signal Fusion and Semantic Similarity Evaluation for Social Media Based Adverse Drug Event Detection," *MS Thesis, Dept. of Comp. Sci. & Elec. Engg., West Virginia University*, Morgantown, WV, USA, 2018.
17. G. A. Miller, "WordNet: a lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.
18. M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, "The SIDER database of drugs and side effects," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1075–D1079, 2016.