SIGNAL MINING IN FINE-GRAINED SENTIMENT ANALYSIS

Christy P. Wong, Marshall School of Business, University of Southern California,
Los Angeles, CA 90089, Christy.Wong.2018@marshall.usc.edu
Wang-chan Wong, College of Business and Public Administration, California State University,
Dominguez Hills, Carson, CA 90747, wcwong@csudh.edu

ABSTRACT

Sentiment analysis is important to the understanding of online purchasing behavior. A fine-grained sentiment analysis is used to identify the sentiments of features (attributes, properties, or aspects) of a product that are extracted from the reviews. A feature-based sentiment analysis can offer in-depth insights of the consumer's perception of a product. Sentiments of features change over time; if a feature change is out of the ordinary, it becomes a signal that may require further investigation. In this paper, we propose a novel signal mining approach adopted from disproportionality analysis in pharmacovigilance. The proposed approach supports signal mining in cohort analysis, and offers a promising new method of mining online reviews.

Keywords: Signal mining, Fine-grained sentiment analysis, Disproportionality analysis, Cohort analysis

OVERVIEW

There are numerous studies demonstrating that online reviews will make or break a product [2]. According to the survey by [1], 84% of people trust online reviews as much as personal recommendations. Over 90% of consumers admit that they are heavily influenced by online reviews. Obviously, free format product reviews are more in depth than star rating systems or Like/Dislike rankings; they reveal affection and emotional feeling of the reviewer towards a product [6]. Sentiment analysis (also known as opinion mining) refers to the use of natural language processing (NLP), textual analysis, and computational linguistics to identify and extract the affection and emotional feeling towards a product in customer reviews. The analysis classifies online reviews into categories such as negative, neutral, and positive. To an analyst, knowing these high-level sentiments of a product may not be enough to fully understand the consumers. Recently, many studies suggest that fine-grained sentiment analysis can reveal deeper insights [5], [10]. In a fine-grained sentiment analysis, the features, properties, characteristics, and aspects of a product from its online reviews are extracted, categorized and analyzed. For example, in studying online reviews of a hotel, the features can be the hotel's cleanliness, location, staff, services, foods, amenities, and so on [11]. Essentially, a fine-grained sentiment analysis is to convert unstructured data, i.e. reviews, into structured data, i.e. features, that can be analyzed efficiently. In a fine-grained feature sentiment analysis, a signal indicates the unusual change of sentiment towards a feature of the product at a specific time.

In this paper, we introduce a novel approach on mining signal of product feature sentiment by comparing it to its cohorts. Our approach is based on the disproportionality analysis for post-marketing signal detection for the adverse drug reactions (ADRs) in pharmacovigilance. The proposed approach works well and promises to be an effective tool for mining sentiment signals of online reviews.

BACKGROUND

We conducted a sentiment analysis study in China for a U.S. company, collecting 4+ million online reviews of its product from six e-retailers for a 12-month period. We used PolyAnalytst [8] to carry out the NLP textual analysis. As with any sentiment analysis, a domain specific dictionary was created. The ontological taxonomy, i.e. features and domain specific terms, were indexed. We achieved a 90.87% accuracy of parsing the reviews for the domain we were studying. There were many significant insights discovered through data mining for this project. In this paper, however, we focus on one aspect: did the feature sentiment changes indicate any signals that should trigger further investigation?

SIGNAL MINING IN COHORT ANALYSIS

To offer meaningful insights, we propose mining signals of a product feature by comparing it to its cohort. A cohort is a group of subjects who share a common characteristic of the same period. In this study, the company offered three product lines: premium, mid-range, and "special," all under the same brand. These three product lines shared common features extracted from the textual analysis. To illustrate how our approach worked, we picked a positive sentiment on **Feature A** of these three product lines. Their trend lines are depicted in Figure 1. Line 1 is the premium line, while Line 2 and Line 3 are the mid-range and "special" lines, respectively.

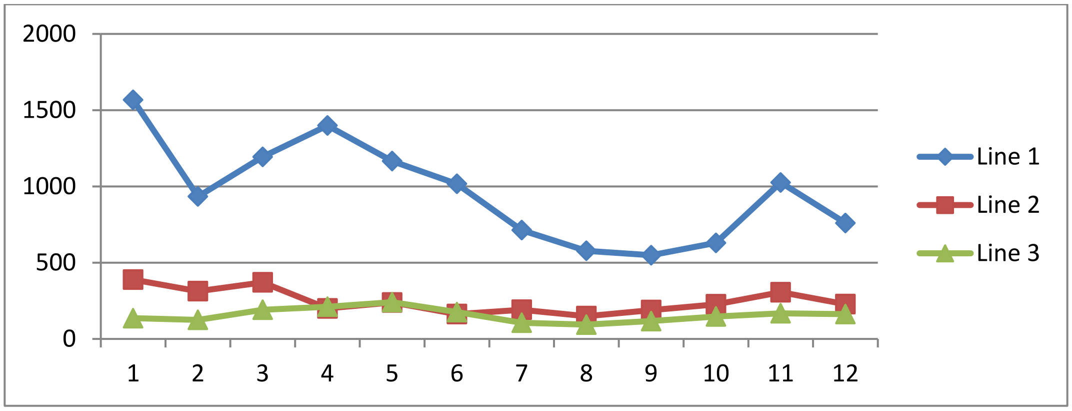

Figure 1 – Positive Sentiment on **Feature A** of the three product lines

Management was eager to find out if there were any signals that would require further investigation. The marketing department was interested in correlating their product enhancement efforts and marketing campaigns to the sentiment changes of **Feature A**.

SIGNAL MINING OF ADVERSE DRUG REACTIONS (ADRs)

Pharmacovigilance is the discipline of monitoring adverse reactions to drugs after they have been in the market place. When a patient takes a drug, and has adverse effects of any sort, the patient and/or the care provider and practitioner voluntarily reports the case to a centralized system such as the US Food and Drug Administration (FDA) Adverse Event Reporting System. The primary purpose of the reporting is to provide early warnings of adverse drug reactions that have not yet been discovered during clinical trials, typically due to limitations of sample size, duration of trails, and so on. The process to identify the unusual ADRs is called *signal mining*. Once a signal is identified, other sources of evidence can be collected and investigated further. If there is sufficient evidence to support the ADRs, interventions will be introduced to minimize the risks and to inform future patients about the drug [3]. There are several techniques used to identify ADR signals. In this paper, we focus on two common techniques that are based on disproportionality analysis techniques: Proportional Reporting Ratio (PRR) and Reporting Odds Ratio (ROR). To understand these two techniques, let's consider a 2 × 2 contingency table as shown in Table 1.

	Specific AE (y)	All Other AEs	Total
Specific Drug (x)	a = 20	b = 100	(a+b)= 120
All Other Drugs	c = 100	d = 980	(c+d) =1080
Total	(a+c)=120	(b+d)=1080	(a+b+c+d)=1200

Table 1 – A 2×2 Contingency Table for ADRs

The table is constructed by summarizing the ADRs from a pharmacovigilance spontaneous report system (SRS) for a specific period. In this example, the reported number of a specific adverse event (AE (y)) of a specific drug (x) is 20 (a) while the reported number of the same drug (x) of all other AEs other than (y) are 100 (b). Similarly, the reported number of AE (y) for all other drugs other than drug (x) is 100 (c) and the reported number of all other AEs for all other drugs is 980 (d). Then PRR and ROR are defined as shown in Table 2:

Definition	Probabilistic Meaning
$PRR = \frac{a/(a+b)}{\mathrm{c}/(\mathrm{c}+\mathrm{d})}$	$\frac{\Pr(AE\ y\vert\ drug\ x)}{\Pr(AEy\vert\neg drug\ x)}$
$ROR = \frac{a/c}{\mathrm{b}/\mathrm{d}}$	$\frac{\Pr(AE\ y\vert\ drug\ x)/\Pr(AE\ y\vert\neg drug\ x)}{\Pr(\neg AE\ y\vert drug\ x)\ /\Pr(\neg\mathrm{AE\ y}\vert\neg\mathrm{drug\ x})}$

Table 2 – PRR and ROR Definitions

Intuitively, PRR is the probability for a given drug, i.e. drug (x), the proportion of ADRs with a pre-specified AE, i.e. AE(y), among all reports related to that drug divided by the proportion of reports of that same AE among all reports related to all other drugs for the same period. Consequently, a drug that has more of the adverse event reported than other drugs will have a higher PRR value. An ROR is the odds for event in group 1 (AE (y)) divided by the odds for event in group 2 (¬AE y). In addition to the PRR and ROR, the Chi-square of the 2×2 contingency table is computed to measure its association. The null hypothesis $\boldsymbol{H_0}$ assumes that there is no association between the row and column variables while the alternative hypothesis $\boldsymbol{H_a}$ claims that there are associations between them.

Choice of Precision Estimate and Threshold

PRR and ROR indicate the extent to which an adverse event is reported by patients taking a specific drug, compared to the frequency at which the same adverse event by patients taking other drugs. If the ratio is greater than 1, it suggests that the adverse event has been reported more frequently, relative to the comparison drugs (the cohort). For Table 1, the PRR, ROR and Chi-square are 1.80, 1.96, and 6.58, respectively. For a 2×2 contingency table with df = 1, the computed Chi-square is large enough to reject the null hypothesis (Chi-square values greater than 3.841 indicate statistical significance with $p \leq 0.05$ for df=1). Since the Chi-square is greater than 3.841, it implies that there is a positive association between drug (x) and AE(y). While there is no golden standard on the threshold to identify whether an AE is indeed a signal, a common practice in pharmacovigilance suggests that an AE is a potential signal if PRR and ROR are >=2, Chi-square is >=4 (95% level df=1), and number of reports >=0.2% of total reports [9]. There are three possible consequences after detecting a signal: the signal can be ignored, the incident can be investigated further, or it can initiate some intervention if the adverse event is critical.

ADOPTING DISPROPORTIONALITY ANALYSIS TO SENTIMENT ANALYSIS

In this section, we demonstrate how we adopt disproportionality analysis to feature sentiment analysis. Consider the trend lines in Figure 1 of the three product lines. We define PRR and ROR of the cohort study in Table 3 below.

	Month$_j$	Sum of all remaining months	Total
Line$_i$	a	b	(a+b)
Sum of all other Lines	c	d	(c+d)
Total	(a+c)	(b+d)	(a+b+c+d)

a = frequency of positive sentiment of Month$_j$ of Line$_i$
b= frequency of positive sentiment of all remaining months of Line$_i$
c= frequency of positive sentiment of all other lines of Month$_j$
d= frequency of positive sentiment of all other lines of all remaining months

$$PRR = \frac{a/(a+b)}{\mathrm{c}/(\mathrm{c}+\mathrm{d})} \qquad ROR = \frac{a/c}{\mathrm{b}/\mathrm{d}}$$

Table 3 – Positive Sentiment of ***Feature A*** for Month$_j$

The PRR is defined as the ratio between the frequency of positive sentiment of Line$_i$ of Month$_j$ and the frequency with which the positive sentiments in the comparison group (relative to all other lines and all other months in the cohort group). In this example, the cohorts are consumers who purchased the three product lines and gave their reviews on **Feature A** in a period of 12 months. Table 4 summarizes the frequency of positive sentiments of **Feature A** as depicted by the trend lines in Figure 4.

Month	1	2	3	4	5	6	7	8	9	10	11	12	Total
Line 1	1567	934	1194	1399	1166	1017	713	578	549	630	1026	760	11533
Line 2	389	313	371	200	239	163	191	149	189	227	306	228	2965
Line 3	136	125	191	211	241	176	106	94	117	147	168	163	1875
Total	2092	1372	1756	1810	1646	1356	1010	821	855	1004	1500	1151	16373

Table 4 – Positive Sentiment on ***Feature A*** by Product Line by Month

The analysis calls for a total of **36** 2×2 contingency tables (3 product lines × 12 months.) For instance, the contingency table of Line 1 and January and its computations are summarized in Table 5.

	January	All other months	Total
Line 1	1567 (1473.59)	9966 (10059.41)	11533
All Other Lines	525 (618.41)	4315 (4221.59)	4840
Total	2092	14281	16373

*Numbers in parenthesis are the expected values

PRR = 1.25, ROR = 1.29, Chi-square = 22.97

Table 5 – The 2X2 Contingency Table for Line 1 in January

Table 5 shows that the frequency of positive sentiment of **Feature A** of Line 1 in January is significant since the Chi-square is greater than 4, with an observed frequency of 1567 instead of the expected 1473.59. Both PRR and ROR are greater than 1. Hence, the positive sentiment of **Feature A** of Line 1 in January is a potential signal.

Computational Results

We coded two ***R*** programs to conduct the analysis. First, we used PhViD, an ***R*** package for pharmacovigilance signal detection [7], to analyze the data set. We also coded an ***R*** program to compute the PRR, ROR and the contingency table Chi-squares. PhViD does not compute the Chi-square. Instead, it computes the p-value to identify the significance of the event. With the same data set, both programs identified identically12 signals as shown in Table 6.

Line	Month	Frequency	Exp Value	PRR	ROR	Chi Sq	p-value
1	Apr	1399	1274.95	1.428496	1.487650	45.90632	0.00000
1	Jan	1567	1473.59	1.252601	1.292319	22.96644	0.00000
2	Feb	313	248.46	1.336558	1.376280	22.34732	0.00000
3	May	241	188.50	1.326318	1.374446	18.36162	0.00001
1	Jun	1017	955.15	1.258996	1.284043	14.76956	0.00007
2	Oct	227	181.82	1.321126	1.347749	14.60837	0.00006
2	Mar	371	318.00	1.211333	1.241558	12.08479	0.00024
3	Oct	147	114.98	1.326305	1.354063	10.73126	0.00050
2	Sep	189	154.83	1.283296	1.302584	9.714418	0.00090
3	Dec	163	131.81	1.275667	1.301914	8.965187	0.00131
2	Nov	306	271.64	1.158928	1.177217	5.843968	0.00761
3	Sep	117	97.91	1.225847	1.240878	4.433651	0.01736

Table 6 - 12 Potential Signals Identified by Chi-Squares and by PhViD

Figure 2 depicts these 12 signals on the trend lines.

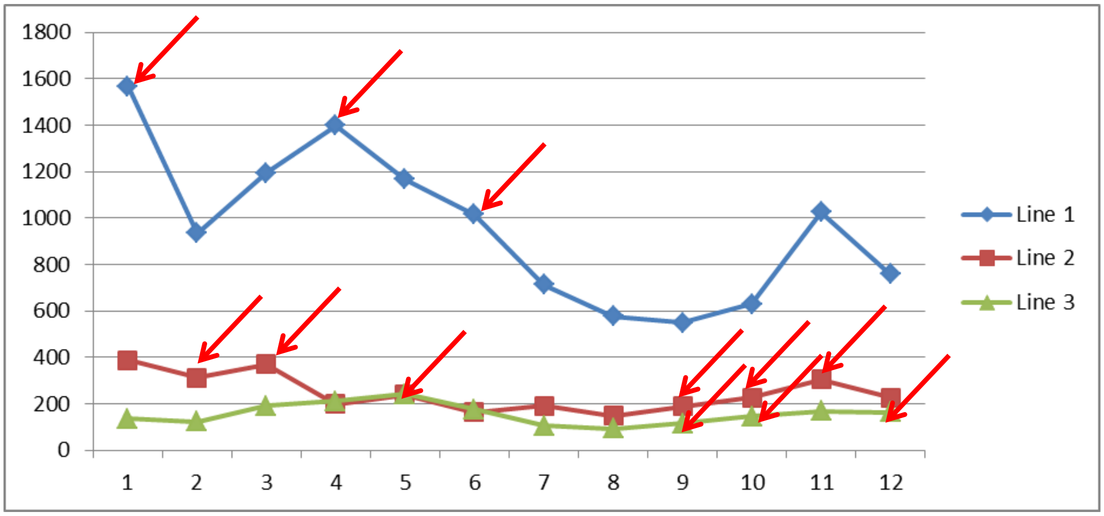

Figure 2 – Signals on the Trend Lines

It is interesting to note that signal may not always be at a rising peak. Consider the signal of Line 2 in February. The PRR looks at the disproportionality at two levels: (1) self-comparison, how its frequency in February is compared to its frequency of the rest of the year, and (2) cohort-comparison, how the frequency of its cohorts in February is compared to the frequency of the cohorts of the rest of the year. In this case, the frequency of Line 2 in February is relatively higher when compared to itself throughout the year and its cohorts. i.e. (Line 1, February) and (Line 3, February). Furthermore, visualizing the scattered signals may offer further insights to the analyst, for example, the cluster of signals from September to December of Line 2 and Line 3.

Like pharmacovigilance ADR surveillance, once signals are identified, it is the beginning for further mining and investigation. The signals being discovered in this approach should offer more insights and directions for these future mining efforts. For instance, further studies discovered that during the months of September to December, the weather was cold. The cold weather benefited Lines 2 and 3 on ***Feature A*** more so than that of Line 1. Furthermore, the marketing department could also relate the results to their campaigns. Consider signal at (Line 1, April). The marketing department conducted a comprehensive campaign to promote **Feature A** of Line 1 in March. The increase of positive reviews received in April could be attributed to the success of the March marketing campaign.

Choice of Threshold

If PRR and ROR are >=1 and the Chi-square of the contingency table is >=4, the observed event could be a signal. Just like in the pharmacovigilance spontaneous ADRs studies, there is no golden standard defining the thresholds of PRR and ROR. For ADRs, the suggested threshold is for PRR and ROR >=2. A high PRR will reduce the number of potential signals. The threshold was first suggested by [3] who observed that if a PRR value close to or less than one may be the consequence of background noise and does not truly represent a signal. In some cases, PRR > 3 was suggested to be the threshold [3]. PRR can also be set up based on risk management plans. Consider the situation if ***Feature A*** could cause an adverse effect, like allergy. A risk management plan (RMP) could determine how extensively an analyst should investigate the signals, thus defining the PRR threshold. Studying the impact of RMP on determining the PRR threshold, however, is beyond the scope of this paper.

CONCLUDING REMARKS

In this paper, we introduced a novel disproportionality analytic technique on fine-grained features of sentiment analysis. Our approach is based on the pharmacovigilance spontaneous ADRs reporting research. The proposed approach offers several unique aspects on sentiment analysis:

1. Our approach focuses on fine-grained feature sentiments. The signals detected will guide further mining efforts. It is analogous to a focused search algorithm in knowledge discovery.
2. Our signal detection is cohort-based. As seen in the above examples, we are not detecting signals of a feature of a product alone. We are detecting signals of a product feature relatively to the same features of the cohorts in the same period.
3. Our approach also offers flexible temporal granularity. Data points can be by days and weeks, instead of months.
4. Conceivably, our approach can incorporate risk management by setting up different thresholds for PRR and ROR.

Limitations and Future Research

The proposed approach is not for general signal mining. Rather, it is for signal mining among cohorts. The approach is *a priori*: the analyst needs to know exactly what to look for and whom to compare with. The second issue is the temporal granularity, that is, the duration and frequency of the measures. If the granularity is changed to finer grained, it will have impacts on the computation. A naïve approach of computation will have the time complexity of $O(mn)$, where m is the growth of the number of cohorts and n is the time unit. Additionally, more data points could potentially discover more signals. This must be approached cautiously. The increase of signal might grow out of control if the PRR and ROR threshold is not well defined. For future studies, we will need to find a way to define the threshold, e.g. through some machine learning models. Or, it could be based on a risk management plan.

REFERENCES

References are available upon request from the authors.