Detecting Adverse Drug Reactions Using a Sentiment Classification Framework

Hashim Sharif and Fareed Zaffar
Department of Computer Science
LUMS University
Email: hashim.sharif,
fareed.zaffar@lums.edu.pk

Ahmed Abbasi
Predictive Analytics Lab
University of Virginia
Email: abbasi@comm.virginia.edu

David Zimbra
Department of Operations & MIS
Santa Clara University
Email: dzimbra@scu.edu

ABSTRACT

Medical blogs and forums are a source of sentiment oriented content that is used in diverse applications including post-marketing drug surveillance, competitive intelligence and the assessment of health-related opinions and sentiments for detecting adverse drug reactions. However applying existing tools for sentiment analysis to health-related datasets provides inadequate classification accuracy. These methods employ less useful features sets and therefore lack discriminatory potential. In this study we propose a framework that uses feature set ensembles with novel feature representations that reduce sparsity by adding representational richness. Our framework extracts important semantic, sentiment, and affect cues, that are better able to reflect the experiences of people when they discuss adverse drug reactions as well as the severity and the emotional impact of their experiences. Experiments conducted on a test bed of health-2.0 datasets, demonstrate improved classification accuracy in comparison to existing techniques. Furthermore, the proposed framework is able to detect adverse drug events earlier, and with higher recall than comparison methods, thereby demonstrating its utility for social media based post-marketing drug surveillance.

I INTRODUCTION

Several key stake-holders in the pharmceutical industry including patients, physicians, regulatory authorities and pharmceutical companies are increasingly using web technologies such as social media, blogs, forums, podcasts and wikis etc to generate and access medical content. Health 2.0, as this is referred to, serves as an important source of on-line medical opinions, information and sentiments relating to particular drugs and events[23][24]. As a result, a patient can make informed decisions about drugs, diseases, procedures and health-care providers.

Social media analytics is a powerful tool that have been used extensively to help mine social media content across several domains[16][19]. More recently, the techniques have been applied to user generated online health and medical content to help guage patient concerns, sentiment and emotions about particular brands, drugs or procedures. Drug related posts in medical forums, are mostly conversational in nature, and are thus representative of prevailing public opinion. This allows stake-holders such as pharmceutical companies to help identify issues faster, compare the online reputation of brands and competitors, post marketing drug surveillance [15][16][17], and the assessment of health related content to predict adverse drug reactions. Importantly, the early prediction of adverse drug reactions is necessary for risk management purposes; single action lawsuits and reputation damage can result in financial losses for the pharmaceutical companies [16]. Accordingly, much research has focused on classifying positive and negative sentiments in on-line medical media, in order to reflect the public inclination towards various drugs.

Existing sentiment analysis tools provide inadequate prediction accuracy, when applied to drug-related posts. The baseline methods employ classifier models trained on words and parts of speech features [22]. Classifier models trained merely on these feature sets cannot benefit from the underlying sentiments, semantics and affects that play a critical role in identifying sentiment polarity. Other problems include, sparse feature vectors[18], representational richness issues, colloquial style of expression [18], and inherent ambiguity with respect to sentiment polarities in medical social media. In this paper we propose a feature representational richness framework (FRRF) for sentiments analysis on Health 2.0 data. Our framework incorporates novel feature representations that extract important sentiments, semantics, affects, and domain specific features. These feature sets are able to communicate important information related to what people are experiencing, and the severity and the emotional impact of the experience. FRRF employs a feature set ensemble approach, that uses a number of parametric feature combinations. Our scheme com-

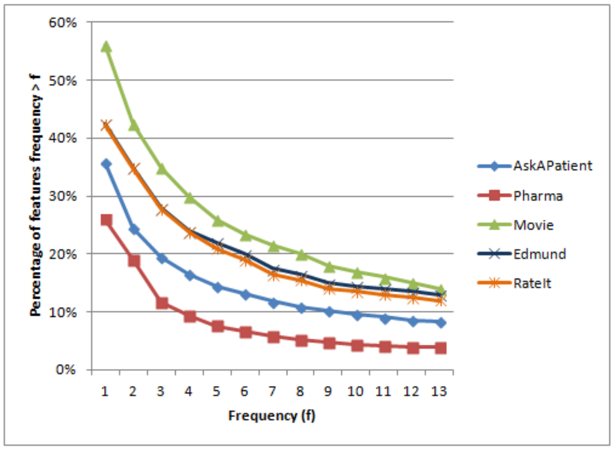


Figure 1: Frequency of features in different datasets

bines the different feature sets to alleviate sparsity and representational richness issues. The feature ensemble models allow for improved sentiment polarity detection in comparison to the individual constituent models. This however, results in a large number of features that add noise and redundancy in the feature space, thereby increasing the likelihood of over fitting. A feature subsumption phase is applied to efficiently remove redundant or less useful n-grams, allowing for more effective n-gram feature sets. Extensive experiments on AskaPatient forum posts and a Twitter dataset of pharmaceutical drugs, reveal that our framework outperforms existing techniques in terms of accuracy and balanced predictions across the sentiment classes. The results have important implications for social media analytics where an accurate prediction framework is critical for the derivation of social intelligence. The remainder of the paper is organized as follows: Section II presents prior relevant work on sentiment analysis and emphasizes the research gaps. Section III describes the proposed feature representational richness framework that demonstrates improved sentiment prediction for medical content, Section IV describes our evaluation test bed and experiments, Section V offers concluding remarks

II RELATED WORK

Many research studies have performed sentiment analysis on online medical data. A similar study, analyses the emotional impact of participation in online health forums on cancer patients [25]. This research employs a computational approach using machine learning and text analytics, to estimate the sentiment of forum posts and discover the sentiment change patterns in thread originators, when they receive social support from other forum members.

Sentiment analysis tools to determine sentiment polarities in opinionated text, can be divided into two major categories; stand-alone and trained workbench tools. Stand-alone tools use text analytics models to label unseen test data immediately; without requiring to train classifier models. SentiStrength [2] a popular stand-alone sentiment analysis tool uses a sentiment lexicon for assigning scores to negative and positive phrases in text. Phrase level scores are aggregated to determine sentence level polarities. Such an approach being unsupervised is easier to apply directly to test data. However the lack of indirect indicators of sentiment and domain specific knowledge degrades classification performance [2]. Workbench tools, on the other hand, employ a supervised learning approach to generate classifier models, trained on labelled data consisting of words and parts of speech tags. The approach requires extensive training and parameter tuning, but possesses the ability to incorporate domain specific features that can serve as indirect indicators of sentiment polarity. The tools however, yield inadequate performance on medical datasets. While this can mostly be attributed to sparsity in feature vectors, health-2.0 content also embodies a bulk of noise features. Furthermore challenges include irrelevant content and the conversational style adopted by typical users[18]. We evaluated the performance of the proposed feature representational richness framework and five popular freely available tools on two different medical datasets. The first dataset, comprises of over 114K forum posts derived from the AskaPatient medical forum. The second dataset, Pharma is a collection of 5K tweets, pertaining to pharmaceutical drugs. The evaluated comparison tools include SentiStrength [2], Sentiment140 [3], OpinionFinder [10], FSH [1][14] and a word n-gram baseline [22]. The factors that lead to diminished performance for the existing tools, are described in the following paragraphs; they also provide the key design features for our proposed framework.

A: Challenges in Online Medical Sentiment Analysis

a) The users of blogs and forums use colloquial language for the purpose of convenience. Furthermore, most comments posted on the medical forums are short, with limited sentiment cues. This leads to feature sparsity and representational richness issues for text based analytics; classifiers require dense feature vectors in order to train a high performing prediction model. [Figure 1] illustrates the feature sparsity

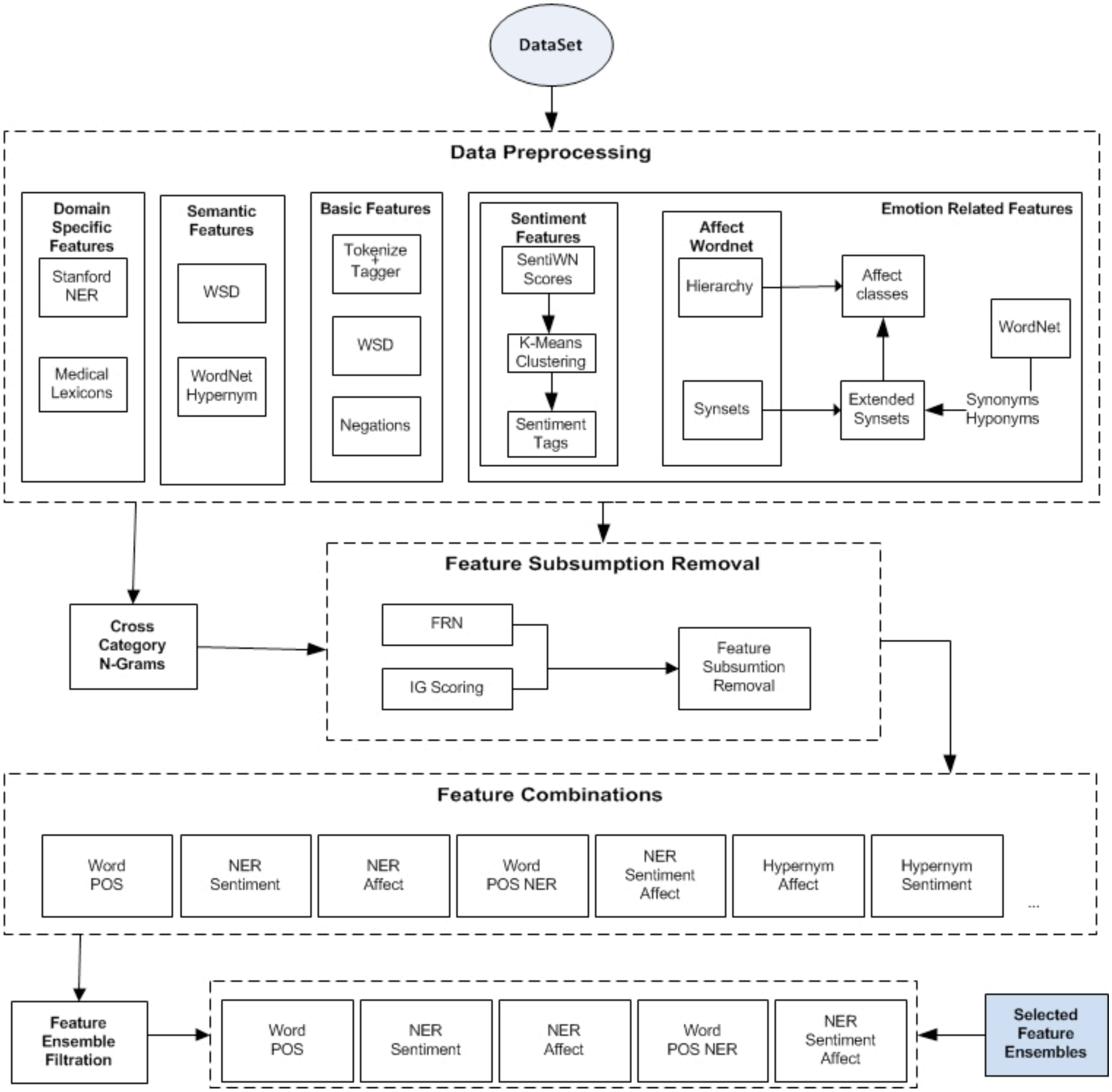


Figure 2. Schematic diagram of FRRF

problem in the medical datasets. The figure shows the unigram occurrence frequencies for three product review datasets (RateIt, Edmund [1], RottenTomatoes [19]) and two of the medical datasets employed in our evaluation (Pharma, AskaPatient). Only 16-25% of the features occur more than twice in the medical forums, as compared to product reviews where 35- 45% occur more than twice. The product review datasets include 100 words per instance, the AskaPatient dataset contains 30 words per instance and the Twitter Pharma dataset contains merely 20 words per instance [18]. As a consequence of applying frequency thresholds, sparse testing and training matrices are generated that in turn lead to diminished classification accuracy. There is a need for incorporating more generalizable lexicon-based knowledge into the sentiment classifiers in order to assist in the extraction of common underlying semantics, sentiments and moods.

b) Health-2.0 content contains a bulk of domain specific information including drugs names, reactions and other medical related entities. Such information is instrumental in determining the polarity of text. This necessitates the need for incorporating medical domain specific features into the model.

III FRAMEWORK DESIGN

In this section, we propose a feature representational richness framework (FRRF) that incorporates novel feature representations that are able to extract useful underlying indicators of sentiment polarity in order to address the challenges highlighted in the previous section. At a higher level, the schematics of our framework [Figure 2] can be visualized as a multilayer stack [9]. The first phase in the process involves the generation of diverse feature representations. The feature subsumption removal phase leverages the predefined relationships between different feature types within and across representations, to remove the redundant and less useful features [1]. The constructed feature sets are used to create feature parametric ensemble models; whereas a combination of different feature sets is used to train classifier models. Moreover, feature parametric model selection is applied to filter out the low performing models. The selected parametric models are used to train classifiers using a Support Vector Machine (SVM). The added feature representations reduce sparsity in feature vectors by extracting important semantic, emotion, and domain specific features, consequently improving polarity classification. Additionally, balanced recall rates are attained across the positive, negative and neutral sentiment classes.

The feature sets can be broadly categorized into four feature categories namely baseline features, semantic features, emotion related features and domain specific features. For each of these representations, unigrams, bigrams, trigrams and cross category n-grams are included. A description of the implemented cross category n-grams will be detailed in a later section. Each of the components in the framework are discussed in the next section.

1) Baseline Features FRRF uses word n-grams along with parts of speech n-grams as the primary baseline features. Furthermore, in our background analysis, considering the specific meanings of words by adding word-sense disambiguation features better representation of nouns, verbs, and adjectives. Moreover, negations and booster words have an important impact on sentiment polarity [2]. Accordingly, FRRF includes provisions for handling negations and booster words based on predefined lexicons comprising of commonly used negations and booster words. The adopted procedure negates / boosts all words between the negation/booster word up until the next punctuation [22].

```
Input: Representations = {word, POS, word_POS,
       word_sense, semantic, NER, sentiment, affect}
       each itself an array of size $n$
Output: enhanced feature set
$Features \leftarrow \emptyset$;
for $i \leftarrow 1$ to $Representations.size$ do
    for $j \leftarrow 1$ to $Representations.size$ do
        if $i \neq j$ then
            for $k \leftarrow 1$ to $n$-1 do
                $CrossCategory \leftarrow$
                $Representations$[i][$k$] + $Representations$[j]
                [$k + 1$];
                $Features \leftarrow$Features + $CrossCategory$;
            end
        end
    end
    return $Features$;
end
```

Algorithm: Cross Category bi-grams

2) Semantic Features Semantic Features refer to logical entities that group together a number of semantically similar keywords. We incorporate such features to alleviate sparsity, which is quite pervasive in health 2.0 content. The WordNet [4] lexical database serves as a useful resource for mapping words to their respective semantic classes. The hypernym hierarchy in WordNet is established on the basis of a "type of" relationship. FRRF uses the hypernym tree, for labeling semantically similar words with a common semantic class [21]. The semantic class is essentially a word hypernym at a pre-specified level in the tree. Incorporating hypernym tags for health-2.0 content adds features at an appropriate level of generalization; not as specific as word n-grams and not as general as parts of speech tags. Consider for instance, the word unigrams "epilepsy" and "eclampsia", that possess a common hypernym "disorder". While the word unigrams might get filtered out due to cutoff frequency thresholds, the common semantic feature "disorder" is more likely to be retained as it refers to a collection of related disorders, and thereby occurs more frequently in the corpus.

3) Emotion Related Features Subjective words that represent emotions and moods are important indicators of sentiment polarity [20]. These words are often employed by social media users for conveying their opinions regarding specific drugs and their reactions. Therefore mining emotion related features helps identify drugs with potential adverse reactions.

FRRF uses two parallel representations for emotion related features, that are described in the next subsections.

3.1) Sentiment SentiWordNet [5] is a lexical resource for opinion mining that assigns sentiment scores to each synset in WordNet [4]. An overall sentiment score is assigned to each synset by aggregating the scores for positivity and negativity. In order to deduce the underlying sentiments in online medical content, the proposed approach generates sentiment n-grams. The employed sentiment tags are of the form positive#X, negative#X and #neutral. The words with an aggregate sentiment below a specified threshold, are tagged #neutral; whereas the words with an overall positive/negative sentiment score are labeled positive#X and negative#X, respectively. The label ”X” is a positive integer that is directly proportional to the sentiment intensity; and is assigned by mapping the sentiment scores to discrete score intervals. The K means clustering algorithm is applied separately to the negative and positive scores in order to cluster the continuous scores into discrete intervals. It is worth pointing out that SentiWordNet assigns a different sentiment score to each sense of a word. Therefore, FRRF leverages the word sense information generated by the Word Sense Disambiguation module to assign sentiment scores specific to word senses.

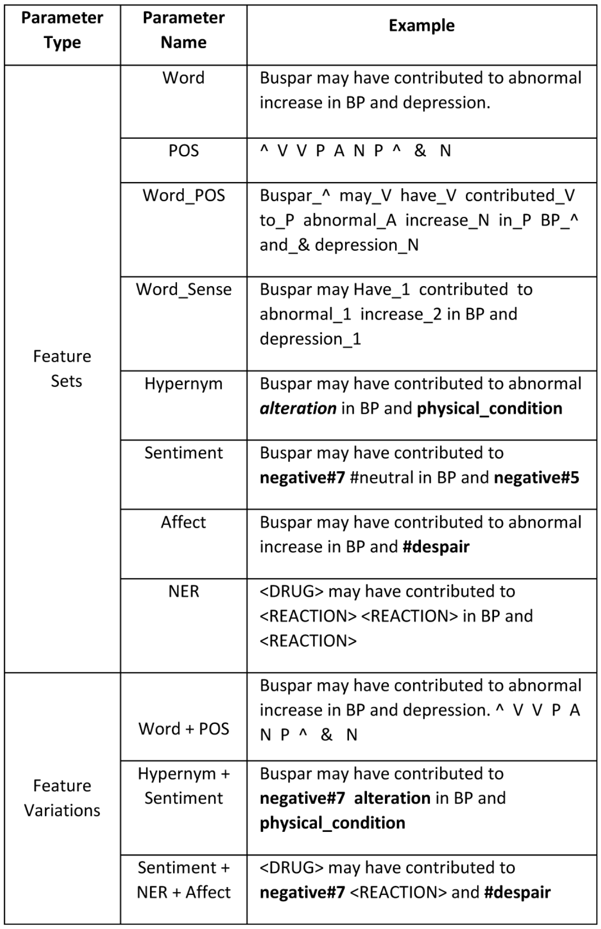
3.2) Affect Affects refer to the emotions and moods present in directional text. Whereas sentiment tags corresponding to particular words and phrases are generally mutually exclusive, words can be associated with multiple affects, with varying degrees of intensity [20]. In health-2.0 content affects play a key role in conveying public perceptions and moods [20], and therefore serve as useful features, for health 2.0 analytics. Accordingly, FRRF leverages the underlying emotions pertaining to drug side effects and related events by incorporating affect tags as a parallel feature representation. AffectWordNet [6] is a lexical resource which includes a subset of WordNet synsets that represent affective concepts. Furthermore, the affect synsets are hierarchically organized into a tree structure. Our extraction approach selects the affects at a pre-specified level in the affect hierarchy as affective classes. The affect synsets in the subtrees rooted at the affective class nodes, are assigned to their corresponding class. In order to further extend the affect synsets in each class, the framework leverages the lexical relations of hyponymy and synonymy between affect synsets in AffectWordNet and related synsets in WordNet. FRRF employs a total of 31 affective classes, some of which include ”despair”, ”sadness”, ”fear”, ”joy”, ”liking”, and ”affection”.

```
Input: Features[], crossCategoryFeatures[],
Output: Removing subsumed cross category features
for $i \leftarrow 1$ to $crossCategoryFeatures.size$ do
    $crossCategory \leftarrow crossCategoryFeatures[i]$;
    $firstFeature \leftarrow getFirstFeature(crossCategory)$;
    $secondFeature \leftarrow$
    $getSecondFeature(crossCategory)$;
    $firstIG \leftarrow getInfoGain(firstFeature)$;
    $secondIG \leftarrow getInfoGain(secondFeature)$;
    $complexIG \leftarrow getInfoGain(crossCategory)$;
    if $firstIG > complexIG$ $OR$ $secondIG > complexIG$
    then
        $Features \leftarrow Features - crossCategory$;
    end
end
return $Features$;
```

Algorithm: Feature Subsumption Removal

4) Domain Specific Features Online medical content contains a plethora of domain specific knowledge pertaining to drugs, reactions, anatomy etc. In addition to reducing feature sparsity, exploiting domain specific features in sentiment classifiers helps generate feature patterns with high discriminatory potential. Existing sentiment analysis tools employ named entity recognition systems, to label names of people, organizations and locations. However, for health-2.0 content there is a need for detecting named entities specific to the medical domain. Consequently, the presented framework includes provisions for labeling named medical entities using medical lexicons. The proposed framework comprises of a name entity recognition (NER) module that uses name lexicons for ”Drugs”, ”Anatomy”, ”Reactions” and ”Administration”. The descriptions for each lexicon is illustrated as follows:

- The ”Drugs” list is a collection of over 4000 common drug names,
- The major internal and external human body parts are listed in the ”Anatomy” lexicon,
- The medical administrative entities, are enumerated in the ”Administration” lexicon,
- The ”Reactions” lexicon includes frequently discussed reactions including ’amnesia’, ’angina’ among others.



Parameter Type	Parameter Name	Example
Feature Sets	Word	Buspar may have contributed to abnormal increase in BP and depression.
	POS	^ V V P A N P ^ & N
	Word_POS	Buspar_^ may_V have_V contributed_V to_P abnormal_A increase_N in_P BP_^ and_& depression_N
	Word_Sense	Buspar may Have_1 contributed to abnormal_1 increase_2 in BP and depression_1
	Hypernym	Buspar may have contributed to abnormal ***alteration*** in BP and **physical_condition**
	Sentiment	Buspar may have contributed to **negative#7** #neutral in BP and **negative#5**
	Affect	Buspar may have contributed to abnormal increase in BP and **#despair**
	NER	<DRUG> may have contributed to <REACTION> <REACTION> in BP and <REACTION>
Feature Variations	Word + POS	Buspar may have contributed to abnormal increase in BP and depression. ^ V V P A N P ^ & N
	Hypernym + Sentiment	Buspar may have contributed to **negative#7 alteration** in BP and **physical_condition**
	Sentiment + NER + Affect	<DRUG> may have contributed to **negative#7** <REACTION> and **#despair**

Table 1. Parallel Feature Representations

Importantly the NER module is not limited to medical related entities, but also incorporates the conventional named entities used in publicly available NER systems; these include ”Person”, ”Location” and ”Organization”. The inclusion of these additional named tags assists the generation of more general extraction patterns. Consider for instance, most of the negative sentiment posts include criticism towards the drug manufacturing pharmaceutical companies. That being the case, labeling all pharmaceutical companies in the corpus as ”Organization”, adds a useful indicator of negative sentiment .

5) Cross Category N-grams In addition to conventional feature n-grams, FRRF also includes cross category n-grams. Cross category n-grams are generated by incorporating features from two distinct feature representations. These mixed n-gram features extract more flexible linguistic patterns in comparison to single representation n-grams. Consider for instance the word n-gram ”Tylenol causes amnesia”. Corresponding to the word phrase ”DRUG causes negative#6” is an examples of a generated cross representation n-gram. It can be observed from the given example that the mixed n-gram represents the more general pattern which in turn proves to be a better indicator of negative sentiment towards a target drug. Hence, utilizing these features can potentially improve the detection of adverse drug reactions. The algorithm for generating cross category bi-grams is presented in [Figure 3]. An array of feature representations is input to the algorithm; whereas each feature representation is in turn an array of features within the representation. It is important to note that the tokens are aligned across the feature representations; with corresponding features across representations occurring at the same indices. The generation of cross category n-grams is similar to conventional bi-grams, apart from the two constituent features belonging to different feature representations.

6) Feature Removal The employed parallel feature representations alleviate the representational richness issues, however the bulk of features generated yield noise and redundancy in the feature space [1]. This prevents quality features from being incorporated due to computational limitations. The removal phase identifies and removes features that are not useful opinion indicators. Importantly a feature relation network (FRN) is employed to define the syntactic relationships between n-gram features. Specifically, only the complex features (bi-grams, tri-grams and cross category n-grams) that outperform their constituent unigrams are retained.

7) Feature Parametric Combinations As previously emphasized, sparse feature vectors are a prevalent issue that persists with social media content. In order to create denser feature vectors, there is a need for maximizing the classification potential of the eight parallel feature representations. Thereby, the framework uses combinations of feature sets to create feature parametric models [18]. Each feature representation adds high quality features contributing towards the creation of high performing classification models. The ensembles demonstrate improved prediction accuracy in comparison to their constituent ensembles. The eight feature representations and a few of the feature parametric combinations along with examples are illustrated in [Table I].

8) Feature Ensemble selection It is worth mentioning that not all feature ensembles generate useful models for sentiment analysis. Therefore, FRRF comprises of a feature ensemble selection module that filters out ensembles not able to deliver higher prediction accuracy in comparison to its constituent ensembles. For instance the feature ensemble "word + hypernym + POS" is filtered out, if the constituent ensemble "word + POS" provides better accuracy. Utilizing only the most useful feature ensembles avoids unnecessary training and testing overhead, thus improving the runtime of the framework and allowing for timely results.

9) Suitability of Feature Representations Table I illustrates the features generated by each feature representation for a sample post taken from the AskaPatient medical corpus. The example is presented to demonstrate the potential of the added feature representations in extracting sentiment cues. It is important to note that the sample post is misclassified by the Word + Pos baseline. However, Importantly the richer feature ensembles comprising of the sentiment, affect, NER and hypernym feature sets correctly classify the post as having negative sentiment polarity. The post clearly states the discontent of a user with regard to the drug "buspar". However, the baseline representations are not able to gather the underlying emotions and sentiments that form the basis of sentiment directionality. Contrastingly, the richer feature ensembles deduce important indicators of negative sentiment polarity. It can be observed that the negative sentiment tags negative#7 and negative#5 corresponding to the words 'abnormal' and 'depression' respectively, serve as important features for classification. Furthermore the affect tag #despair adds another useful indicator of negative sentiment polarity. Notably, the framework also leverages the medical named entities such as < Drug> < Reaction> to incorporate domain specific knowledge.

IV EXPERIMENTS

The proposed feature representational richness framework has been evaluated on two different medical datasets; AskaPatient and Pharma. The experiments conducted include analyzing the sentiment prediction accuracy for each dataset using the proposed framework against comparison tools, and creating a sentiment Index time series for predicting adverse drug reactions. The AskaPatient dataset is a collection of over 114K forum posts. Pharma is a collection of over 5000 tweets related to pharmaceutical drugs. Pharma is a pre-annotated dataset with three sentiment polarity labels; negative, positive and neutral. The AskaPatient posts are annotated with a user rating from 1 to 5. The dataset is split into two sentiment classes; whereas ratings 1-2 are labeled negative and 4-5 are labeled positive. Posts with rating 3 are ignored, as they cannot be regarded as truly negative, positive nor neutral. Therefore it is important to note that the sentiment classification on the AskaPatient posts is a binary classification problem, as opposed to a tertiary classification problem, as in the case of Pharma. For evaluation on the AskaPatient dataset, we use 24K posts for training (12K negative; 12K positive), and the rest 90K are kept for testing. For a total of 5009 Pharma tweets 1350 are utilized for training (450 for each negative, positive and neutral) and the rest are reserved for testing. The tokenization of words and parts of speech tagging is performed using the CMU POS tagger [12]. Word sense Disambiguation is achieved, using the WSD [13] module developed by Mihalcea et al. The semantic hypernym features are extracted from the hypernym trees in the WordNet [4] lexical database. The framework comprises of a named entity recognition module that employs a pre-trained StanfordNER [7] classifier, in addition to using custom built medical lexicons for drugs, reactions, anatomy and administration. In order to label words to their corresponding sentiment tags, we use the popular sentiment lexicon SentiWordNet [5]. Affect tags are derived from the AffectWordNet [6] lexical resource comprising of WordNet synsets representing affective concepts. For training the feature parametric models, FRRF uses the SVMPerf [8] classifier with a liner kernel. A total of 255 feature parametric models are produced. However the ensemble filtration phase, retains only the 31 highest performing models. For the three class classification problem, each parametric model is a one-against-one classifier comprising of three binary classifiers; negative vs. positive, negative vs. neutral and positive vs. neutral. Feature selection was performed at the binary level using the Information Gain heuristic to rank features. A frequency threshold of five is applied thereby filtering out features occurring lesser than 5 times in the training set.

1) Sentiment Classification Results

[Table II] shows the evaluation results for FRRF and the comparison tools on the forum and Twitter datasets. In addition to the overall accuracy, class level precision and recall are used as evaluation metrics. The experimental results demonstrate that FRRF outperformed the comparison methods with respect to overall accuracy and class level precision. It is also worth noting that recall is balanced across the classes. Importantly, FRRF has shown a 9% accuracy improvement over the highest performing comparison tools OpinionFinder[10] and FSH[1] on the Pharma

	AskaPatient Forum Posts					Pharma tweets						
		% class level recall		% class level precision		% Acc.	%class level recall			%class level precision		
Tool	% Acc.	Neg	Pos	Neg	Pos		Neg	Pos	Neu	Neg	Pos	Neu
FRRF	**78.2**	77.85	79.4	82.01	74.86	**79.73**	51.42	53.96	83.27	14.09	39.33	95.07
FSH	76.24	74.05	78.8	80.91	71.55	70.71	21.04	39.15	84.86	28.4	47.57	78.84
SentiStrength	62.11	88.42	27.9	61.46	64.95	55.29	67.63	17.84	61.37	18.82	87.42	79.06
Sentiment140	57.17	82.38	26.56	57.66	55.39	62.09	62.59	44.03	65.84	21.1	83.25	82.06
OpinionFinder	57.64	68.26	44.62	60.17	53.42	70.73	46.66	23.47	76.32	6.74	33.47	91.11
Word Baseline	71.99	74.43	69.04	74.40	69.08	68.5	51.42	53.26	71.80	12.82	24.02	94.10

Table II. Experimental Results for FRRF and Comparison tools on a test bed of two medical datasets

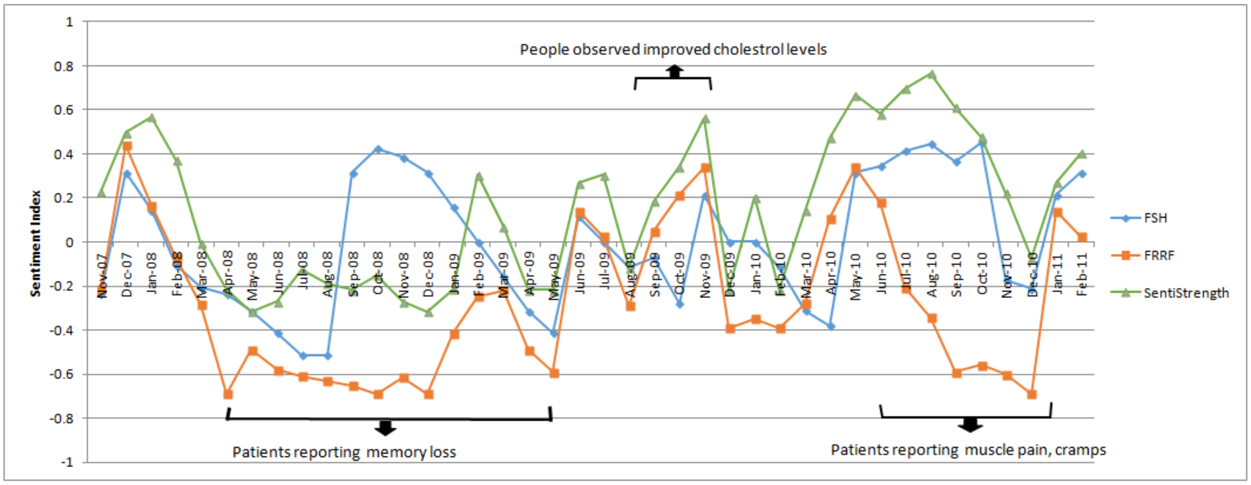


Figure 3. Sentiment Index Time Series for drug 'vytorin'

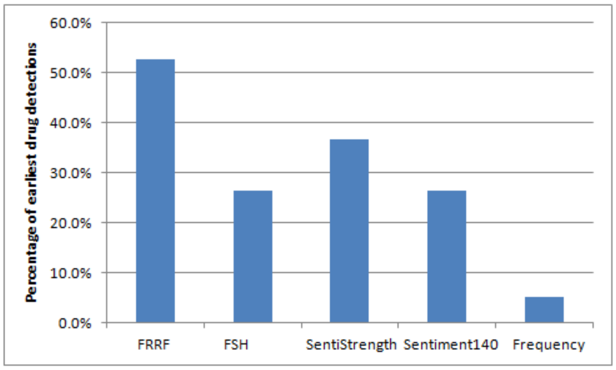


Figure 4. Earliest Drug Predictions across comparison tools

Twitter data Set. The accuracy improvement over FSH[1] on the AskaPatient data Set is 2.5%. The percentage improvement on the Pharma Twitter dataset is more significant owing to the higher sparsity in tweets in comparison to the forum posts. However the improvement is significant for the large forum dataset with a test set of 90K posts. The stand alone tools, SentiStrength [2], Sentiment140 [3] deliver significantly lower performance, primarily due to low positive recall. Notably, by incorporating diverse indicators of sentiment and flexible linguistic features, FRRF is able to outperform the trained work bench tools, FSH [1] and word baseline [22] with regard to negative recall. Improved negative recall and neg-

Tool	Tool Type	Overall recall	Relative Recall	False Positive (%)
FRRF	Workbench	60	70.588	35.32
FSH	Workbench	55	64.70	35.57
SentiStrength	Stand-alone	45	52.94	44.02
Frequency	Baseline	50	58.82	62.93
Sentiment140	Stand-alone	45	52.94	45.36

Table 3. FRRF drug recall versus comparison tools

ative precision are particularly important, as they are representative of better public sentiment detection with regard to potential adverse drug reactions.

2) Sentiment Index Time Series and Adverse Events The results presented in the previous subsection have demonstrated improved sentiment classification performance of FRRF over comparison tools. As previously mentioned, the motivation of improving sentiment classification on medical social media is to improve the detection of potential adverse drug reactions. Users share their opinions and views regarding certain drugs on social media, adding sentiment patterns over time, that can in turn serve as useful indicators of adverse drug reactions. In order to identify these sentiment patterns and correlate them with actual events, we constructed a sentiment time series on 90K posts derived from the AskaPatient data Set. FRRF classifies each post into positive or negative sentiment polarity. The sentiment classifier was trained on a set of 24K posts, used in the experiments detailed in the previous sub-section. The time series presented in[Figure 3] shows the monthly average sentiment polarity for the drug Vytorin (aggregated for all the posts in the month) over a time period ranging from November 2007 to February 2011. It is important to note that a potential adverse reaction of Vytorin was detected by the FDA in June 2011. Important sentiment patterns prevailing over a period of time pertaining to the drug are annotated in the timeline. The x-axis depicts the month and the y-axis shows the sentiment index ranging from -1(extreme negative) to 1 (extreme positive). The figure depicts the sentiment index time series generated by FRRF in comparison to benchmark tools FSH [1] and SentiStrength [2]. As illustrated in the figure, the FRRF time series demonstrates better correlation with actual events in comparison to the benchmark methods. For instance, in the time period from March 2008 to March 2009, a consistently negative sentiment pertaining to Vytorin was observed, mainly due to complaints regarding the side effects. In this time span, many people had been reporting mental confusion, and memory loss arising from an increase in cholesterol levels. Similarly, from mid-2010 to late-2010, a highly negative sentiment regarding the drug was observed. Many people had attributed muscle pain, cramps and tiredness to the use of Vytorin. The example suggests that in addition to better reflecting drug-related sentiment indexes over time, FRRF may be capable of improving adverse drug event detection both in terms of accuracy and timeliness of detections. In [Figure 3] , FRRF is able to detect an adverse reaction regarding Vytorin (muscle damage and myopathy) 4 months before comparison tools.

In order to empirically demonstrate FRFFs enhanced adverse drug event detection capabilities, we analyzed 20 drugs with FDA adverse event reports in 2012. For each drug, a monthly sentiment index time series was constructed across 114K forum postings between 2007 and 2011 using FRFF, FSH, SentiStrength, and Sentiment-140. We also included a basic mention model as a baseline (i.e., a time series of the number of monthly mentions of the drug). For each time series, an adverse event trigger occurred if the negative sentiment z-score exceeded a value of 3 for a given month. The results are presented in [Table 3] and [Figure 4]. FRRF outperformed the comparison tools with respect to overall recall, relative recall, and false positive rates. The overall recall refers to the fraction of correctly identified adverse drugs, present in the dataset. In contrast, relative recall illustrates the fraction of correctly detected adverse drugs from a subset of total drugs; comprising of drugs that were detected by any one of the comparison tools. Moreover, FRRF demonstrated the lowest percentage of false positives (i.e., false alarms for drugs that do not appear in any FDA adverse event reports). [Figure 4] draws a comparison of the various tools, in terms of earlier detection of adverse drug reactions. It can be observed that FRRF exhibits the highest percentage of earliest drug detections, with the earliest detection for over half the drugs (in months). It is important to note that the sum of percentage earliest detections exceed 100% since for some drugs, multiple tools tied for the earliest detection month (resulting in double-counting). Overall, the results demonstrate that the balanced and accurate performance of FRRF facilitates improved classification of online drug-related sentiments compared to other sentiment analysis methods and a mention-based baseline, resulting in enhanced detection of adverse drug reactions using social media.

V CONCLUSION

In this study, we proposed a sentiment classification framework for detecting adverse drug reactions. The framework leverages novel feature representations that extract the underlying sentiments in medical social media content. The added feature representations create features with high discriminatory potential for various sentiment classes. Experiments performed on two medical datasets have shown markedly improved sentiment prediction accuracy. Furthermore, the frame-work generates an accurate sentiment time series that nicely correlates with the prevailing public sentiments regarding various drugs. Consequently, the proposed framework facilitates enhanced detection of adverse drug events, with both better event recall and timelier identification of events. While tested in the context of adverse drug events, the framework is general enough to be applied to datasets in other domains. The results have important implications for predictive analytics and social intelligence.

VI ACKNOWLEDGEMENTS

We would like to thank the US National Science Foundation for their support through grant IIS-1236970.

VII REFERENCES

[1] A. Abbasi, S. France, Z Zhang, H. Chen. “Selecting attributes for sentiment classification using feature relation networks.” Knowledge and Data Engineering, IEEE Transactions on 23.3 (2011): 447-462.

[2] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas “Sentiment strength detection in short informal text.” Journal of the American Society for Information Science and Technology (2010).

[3] A. Go, R. Bhayani, and L. Huang. “Twitter sentiment classification using distant supervision.” CS224 Project Report, Stanford (2009): 1-12.

[4] G. A. Miller. “WordNet: a lexical database for English.” Communications of the ACM 38.11 (1995): 39-41.

[5] S. Baccianella, A. Esuli, and F. Sebastiani. “Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining.” Proceedings of the 7th International Language Resources and Evaluation Conference, May. 2010.

[6] A. Valitutti, C. Strapparava, and O. Stock. “Developing Affective Lexical Resources.” PsychNology Journal 2.1 (2004): 61-83.

[7] B. MacCartney. The Stanford Natural Language Processing Group [Online]. Available:nlp.stanford.edu/software/

[8] T. Joachims. “Training linear SVMs in linear time.” Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006.

[9] A. Abbasi, C. Albrecht, T. Vance, and J. Hansen “MetaFraud: A Metalearning Framework for Detecting Financial Fraud,” MIS Quarterly, 36.4 (2012): 1293-1327

[10] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi and S. Patwardhan. “OpinionFinder: A system for subjectivity analysis.” Proceedings of HLT/EMNLP on Interactive Demonstrations, 2005.

[11] D. Zimbra, A. Abbasi and H. Chen. “A Cyber-Archaeology Approach to Social Movement Research: Framework and Case Study,” Journal of Computer-Mediated Communication, vol. 16,2010.

[12] O. Owoputi, B. OConnor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith (2013).“ Improved part-of-speech tagging for online conversational text with word clusters. In Proceedings of NAACL-HLT.

[13] R. Sinha, and R. Mihalcea. “Unsupervised graph-based word sense disambiguation.” Recent Advances in Natural Language Processing V: Selected Papers from RANLP (2009).

[14] A. Abbasi. “Intelligent Feature Selection for Opinion Classification in Web Forums.” IEEE Intelligent Systems 25.4 (2010).

[15] A. Abbasi, F. M. Zahedi and S. Kaza. “Detecting Fake Medical Web Sites using Recursive Trust Labeling,” ACM Transactions on Information Systems, 30.4 (2012): no. 22.

[16] A. Abbasi, T. Fu, D. Zeng, and D. Adjeroh. “Crawling Credible Online Medical Sentiments for Social Intelligence.” Proceedings of the ASE/IEEE International Conference on Social Computing (2013)

[17] T. Fu, A. Abbasi, D. Zeng and H. Chen. “Sentimental Spidering:Leveraging Opinion Information in Focused Crawlers,” ACM Transactions on Information Systems, 30.4, 2012.

[18] A. Hassan, A. Abbasi, and D. Zeng. “Twitter Sentiment Analysis: A Bootstrap Ensemble Framework.” Social Computing (SocialCom), 2013 International Conference on. IEEE, 2013.

[19] A. Abbasi, H. Chen, and A. Salim. “Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums,” ACM Transactions on Information Systems, 26.3 (2008): no. 12.

[20] A. Abbasi, H. Chen, S. Thoms and T. Fu. “Affect Analysis of Web Forums and Blogs using Correlation Ensembles,” IEEE Transactions on Knowledge and Data Engineering, 20.9 (2008): 1168-1180.

[21] S. Scott and S. Matwin. “Text classification using WordNet hypernyms.” Use of WordNet in natural language processing systems: Proceedings of the conference. 1998.

[22] B. Pang, L. Lee and S. Vaithyanathan. “Thumbs up?: sentiment classification using machine learning techniques.” Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.

[23] B. Hesse, D. Hansen, T. Finholt, S. Munson, W. Kellogg and J. Thomas, “Social Participation in Health 2.0,” IEEE Computer, vol. 43, no. 11, 2010.

[24] G. Eysenbach. “Medicine 2.0: Social Networking, Collaboration, Participation, Apomediation, and Openness,” Journal of Medical Internet Research, vol. 10, no. 3, p. e23, 2008

[25] B. Qiu, K. Zhao, P. Mitra, D. Wu, C. Caragea, J. Yen, G.E. Greer, and K. Portier. “Get online support, feel better–sentiment analysis and dynamics in an online cancer survivor community.” In Privacy, security, risk and trust, 2011 IEEE third international conference on social computing