ANALYSIS OF PSYCHOLINGUISTIC PROCESSES AND TOPICS IN ONLINE AUTISM COMMUNITIES

Thin Nguyen, Dinh Phung and Svetha Venkatesh

Deakin University
Geelong, VIC 3220, Australia
{thin.nguyen,dinh.phung,svetha.venkatesh}@deakin.edu.au

ABSTRACT

Current growth of individuals on the autism spectrum disorder (ASD) requires continuous support and care. With the popularity of social media, online communities of people affected by ASD emerge. This paper presents an analysis of these online communities through understanding aspects that differentiate such communities. In this paper, the aspects given are not expressed in terms of friendship, exchange of information, social support or recreation, but rather with regard to the topics and linguistic styles that people express in their on-line writing. Using data collected unobtrusively from LiveJournal, we analyze posts made by ten autism communities in conjunction with those made by a control group of standard communities. Significant differences have been found between autism and control communities when characterized by latent topics of discussion and psycholinguistic features. Latent topics are found to have greater predictive power than linguistic features when classifying blog posts as either autism or control community. This study suggests that data mining of online blogs has the potential to detect clinically meaningful data. It opens the door to possibilities including sentinel risk surveillance and harnessing the power in diverse large datasets.

Index Terms—autism, psycholinguistic, social media

1. INTRODUCTION

Autism is a developmental disorder that causes deficits in social interaction, communication, and behaviors, and is often known as the triad of impairments. Early intervention plays a crucial role and may help an ASD child to substantially overcome these deficits and lead a normal life. Nonetheless, it is a disorder that needs a continuing network of supports from parents, relatives, friends, neighbors and every caring individual to sustain functionality and to fully integrate in society. It is widely acknowledged that collecting individual data for research purpose poses serious privacy concern. Obtaining data at an individual level might risk being misused [1]. Collecting data and performing group analysis from web-logs sidesteps this issue, providing an unobtrusive approach to data analysis, and is advocated in recent work [2]. Also, the physically isolated and mentally ill are over-represented in social media communities for a number of reasons, including limited mobility or independence, difficulties with face-to-face communication, and the scarcity or scatter of fellow sufferers. In these conditions, social media can be a life saver.

Our modern living is interwoven with social networking, epitomized by the popularity sites such as LiveJournal. People use sites like Tumblr, Facebook, and Google+, to connect and share experiences, find answers to health questions, and express themselves with writing (Blogger), photography (Flickr), video (YouTube), and many other kinds of data about their experiences (patientslikeme.com). Tumblr alone has over 70 million users, more than half under 25 years, and is acknowledged as one of the biggest sites offering support for the mentally ill. Online communities of autism emerge in these social media platforms. These communities involve people who are affected or intrigued by ASD. This paper presents an analysis of these online communities, focusing on psycholinguistic processes and topics expressed in the content of posts written by members of these communities. Distilling these aspects of online autism communities could provide valuable information such as language styles or topics of concerns for this community - people affected by ASD, therapists, autism professionals, researchers, and related bodies.

This study examines a large corpus of blog posts crawled from LiveJournal to analyze the discrimination in language styles and topics expressed by autism people in their online diaries. We note that the such data can be obtained unobtrusively and that the algorithms we construct are privacy preserving. In particular, we assume that people in autism communities are different from others in the topics they choose to write about and the way they write. This assumption is evaluated by means of both conventional statistical tests as well as machine learning algorithms.

The remainder of this paper is organized as follows. Related work is discussed in Section 2. Experimental setup and analysis are explained in Section 3, including results for sta-
tical tests of equality in writing styles and content between autism and control communities as well as the performance of community classification with different features. Section 4 concludes the paper and notes potential implications.

2. RELATED BACKGROUND

There has been an explosion in the scope and use of social media, whereby people divulge large amounts of data in real-time. In parallel, rapid advances in machine learning and data mining have been made [3]. This allows us to discover patterns in data on a theoretical basis, and make predictions based on patterns detected from this data. As an example, using Twitter posts, pandemic (H1N1) 2009 influenza and validated disease activity were tracked. Machine learning was able to construct a predictive model utilising a million influenza-related tweets. The model could predict disease activity in real-time [4]. As an exemplar, mining the blogs of military personnel involved in operation enduring freedom in Iraq, was able to discern the emotional tone of blogs and detect service personnel experiences and emotional reactions to combat exposure [5]. This further highlights the potential applicability of machine learning to psychiatric clinical and research settings [6].

Collections of text documents can be characterized by topics discussed by people in blog sub-communities [7] and tagged media [8]. In addition to this kind of depression markers, in our analysis, the linguistic styles online autism bloggers expressed in their logs are taken into account in autism community detections. In this study, we use the Linguistic Inquiry and Word Count (LIWC) package [9] to extract linguistic styles conveyed in blog posts. This package was used to extract psycholinguistic features of a corpus of web blogs [10, 11]. They found that these features were significantly different between old and young bloggers as well as between influencers and non-influentials. LIWC features were also extracted from essays written by depressed colleague students [12]. The authors found that depressed students used more first person singular pronouns and more negative emotions words in their essays than those students who did not experience any depressive episode. In a study on suicide using LIWC, it was found that suicidal poets used more first person singular pronouns and less first plural pronouns in their writing than did non-suicidal poets [13]. As shown in these studies, linguistic styles can be considered an indicator of mental problems. Further evidences of word usage and expressive writing to physical and mental health is established in [14]. LIWC is also used in a huge range of research areas in sociology and psychology, including status, dominance and social hierarchy, honesty, deception, thinking styles and individual differences [15].

Most related to our present work is [2] who analyze word usage from blogs of 40 individuals with ASD using LIWC and found no statistical difference in the rate of word usage except more variation in the use of social words in blogs written by ASD individuals. However, their analysis focus particularly on the posts produced by ASD people in their individual blogs. Our analysis, on the other hand, targets posts written by online autism communities in general, and at a gross level, we found a statistical difference across psycholinguistic processes categorized in LIWC.

3. ANALYSIS AND EXPERIMENTS

3.1. Method

We collect data from the LiveJournal website, querying for all communities discussing autism topic. This process results in 10,000 posts made by ten autism communities whose biographies are listed in Table 1. About 30 per cent of posts are tagged with one of 132 predefined moods, which can be classified into emotional themes [16]. The earliest community was created in 2001, thus our dataset spans over 10 years with a total of 1,910 active users. We further construct a control dataset. For diversity, we crawled 10,000 posts made by 100 communities within 10 LiveJournal categories: Advice-Support, Creative-Expression, Entertainment-Music, Fandom, Fashion-Style, Food-Travel, Gaming-Technology, Parenting-Pets, Politics-Culture and Television.

We performed analysis of two aspects: psycholinguistic processes and topics derived from the content. For psycholinguistic processes, we use the the LIWC package [9] in which English words are grouped into psycholinguistic processes1. For topics, we employ a popular Bayesian probabilistic modeling tool – the latent Dirichlet allocation (LDA) [17] – to extract the topics from AUTISM and CONTROL groups. LDA is a hierarchical Bayesian mixture models that assigns each word in a document to a topic, where a topic is a distribution over the set of distinct vocabularies found in all documents. Thus, the content of a topic can be roughly interpreted by inspecting the subset of vocabularies corresponding to the highest probabilities in the topic. Since each word in a document is assigned with a topic, each document can be represented by the proportion or mixtures of topics used to generate that document. Due to the size of the data, we have implemented our own version in C# using Gibbs inference detailed in [18] for the inference part. We set the number of topics to 50, run the Gibbs for 5000 samples and use the last Gibbs sample to interpret the results.

We perform nonparametric Wilcoxon tests on the hypothesis of equal medians in psycholinguistic processes and topics between the autism and control groups. To further substantiate the difference in the use of these aspects between autism and control, topics and LIWC are used as features to classify a post into either autism or control. Modern machine learning supervised classifiers broadly aim to build models from training data and assign a class to a new data point. They

differ widely in principle and we use 15 different classifiers in several classification paradigms. Core implementation of the learning algorithms used Weka data mining software [19]. 10-fold cross validations are run on the two feature sets. This essentially averages the results on 10 runs, sequentially using one held-out data fold for testing and other nine folds for training. To evaluate the performance of classification, we use the accuracy and F-measure scores. Given True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), the accuracy and F-measure for the classification are defined as:

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{2}
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{3}
\]

\[
\text{F-measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}
\]

<table>
<thead>
<tr>
<th>p-value</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motion</td>
<td>0.31628</td>
</tr>
<tr>
<td>Home</td>
<td>0.82966</td>
</tr>
<tr>
<td>Religion</td>
<td>0.2735</td>
</tr>
<tr>
<td>Death</td>
<td>0.48579</td>
</tr>
</tbody>
</table>

Table 2: The LIWC categories whose the Wilcoxon test on the hypothesis \( H_0 \) of equal medians between AUTISM and CONTROL is not rejected.

3.2. Analysis of Psycholinguistic Processes

LIWC assigns words into one of four high-level categories – linguistic processes, psychological processes, personal concerns and spoken categories – which are further sub-divided
higher than that of control ones, 252 vs 174, showing autism people are more engaged in the conversation than are people.

Table 3 further displays the results for six psychological processes. The null hypothesis is rejected in all cases. Figure 1 shows the difference in the use of these processes between autism and control groups. As shown in Figure 1a, control people favor using social and cognitive words in their posts while perceptual and relativity words are in the preference of autism people. Generally, there is no clear difference in the percentage of biological and affective words in the posts made by both groups. However, there exists obvious difference in the use of sub-groups of these two processes. For example, on affective processes, autism group use more negative words, including anxiety, anger, and sadness, and less positive words than do control group; on biological process, autism people use more health and less body, sexual, and ingestion words than do control people.

On personal concerns, three of them are not rejected in the Wilcoxon test: home, religion and death. It can be thought that people in the two populations are equally interested in these matters. On the other concerns, work is found in the preference of autism people while control ones are more interested in leisure and money.

On spoken language, as shown in Figure 2, nonfluences (e.g., er, hm, umm) and fillers (e.g., blah, I mean, you know) are found to be used more by autism than by control group. As explained in [15], it can be because autism people are ‘uncertain or insecure about their topic.’ On the other hand, autism people have lower use of assents (e.g., agree, OK, yes), showing a lower ‘group consensus and agreement’ [15].

### 3.3. Analysis of Topics

We perform another nonparametric Wilcoxon test on the hypothesis of equal medians in topics between autism and control groups. The null hypothesis $H_0$ is considered to be rejected at $p \leq .05$. Almost all of topics (49 of 50) are found statistically different in the use between the two groups. Figures 3a and 3b show the most popular 10 discovered topics for the control and autism communities respectively. The subgroups of autism and the communities of control group are listed on the $y$ axis, and the topics extracted from their posts (a cloud visualization of topics on words\(^2\)) are shown on the $x$ axis. The entries indicate degree of usage of a topic by a group. We see that the topics in these two groups are quite unique. Further analysis of the topics reveal interesting insights - autism group focus strongly on autism-related topics, whilst the control group use more generic topics such as computer, food or fashion.

Furthermore, visual inspection from these plots shows much more variation in the types of topics used in autism group compared with control group, suggesting that autism communities tend to use a larger subset of topics in their

---

\(^2\)The size of a word is proportional to its probability in the topic.
### Table 4: Classification performance.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>std</th>
<th>F-measure</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMO</td>
<td>79.72</td>
<td>1.15</td>
<td>79.72</td>
<td>1.15</td>
</tr>
<tr>
<td>Bagging</td>
<td>78.68</td>
<td>1.18</td>
<td>78.67</td>
<td>1.18</td>
</tr>
<tr>
<td>RandomForest</td>
<td>77.33</td>
<td>0.69</td>
<td>77.25</td>
<td>0.68</td>
</tr>
<tr>
<td>JRip</td>
<td>76.04</td>
<td>0.92</td>
<td>75.99</td>
<td>0.90</td>
</tr>
<tr>
<td>PART</td>
<td>75.99</td>
<td>1.67</td>
<td>75.88</td>
<td>1.72</td>
</tr>
</tbody>
</table>

(a) Classification results for LIWC features.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>std</th>
<th>F-measure</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMO</td>
<td>87.86</td>
<td>0.59</td>
<td>87.84</td>
<td>0.60</td>
</tr>
<tr>
<td>Bagging</td>
<td>85.74</td>
<td>0.69</td>
<td>85.74</td>
<td>0.69</td>
</tr>
<tr>
<td>RandomForest</td>
<td>85.09</td>
<td>0.51</td>
<td>85.07</td>
<td>0.51</td>
</tr>
<tr>
<td>PART</td>
<td>84.44</td>
<td>0.97</td>
<td>84.41</td>
<td>0.97</td>
</tr>
<tr>
<td>JRip</td>
<td>83.68</td>
<td>0.83</td>
<td>83.67</td>
<td>0.82</td>
</tr>
</tbody>
</table>

(b) Classification results for topic features.

---

The results for this two-class classification (control vs. autism) using LIWC and topics as features are shown in Table 4a and Table 4b respectively. We observe that prediction is best achieved when using topics as features, achieving an accuracy of 87.86 per cent (by SMO classifier). The LIWC features also perform well, gaining an accuracy of 79.72 per cent (by SMO classifier) in a 2-class prediction (50% for a random guess since the input is balanced). Without the need for a feature selection stage, the results for the predictions using psycholinguistic styles (through LIWC) are comparable to topics, but at a lighter computational cost. This indicates a potential application of this information for the purpose of analyzing the networking properties of social media.
Large numbers of people use online communities to discuss mental health issues, opening up possibilities for new understanding. We aimed to study the characteristics of those who join autism communities in comparison with those who join other online groups. Machine learning and statistical methods were used to discriminate between these two groups using psycholinguistic processes and content topics extracted from the posts generated by members of the two groups. We found that both the written content and writing style differed significantly between the two groups, resulting in good predictive validity in predicting community membership using topics and psycholinguistic features as features. Clear discrimination between writing styles and content, with good predictive power to classify posts is an important step in understanding new social media and its use in mental health. Results in this paper can form the foundation of early warning systems.

5. REFERENCES


