Abstract—An explanation capability is crucial in security-sensitive domains such as medical applications. Although support vector machines (SVMs) have shown superior performance in a range of classification and regression tasks, SVMs like artificial neural networks (ANNs) lack an explanatory capability. There is a significant literature on obtaining human-comprehensible rules from SVMs and ANNs in order to explain “how” a decision was made or “why” a certain result was achieved, and this paper proposes a novel approach for SVM classifiers. The experiments describe a first attempt to generating text summaries for explaining “why” certain text contents are depressing or funny using a decompositional approach. Learned model parameters are analysed to select informative feature components and filtering is applied to select subsets of more relevant components. Explanation components are then used to rank basic concept-constructs called predicates, which are generated and ranked using a commonsense database. The ranked predicates are then used to generate textual explanations of classification results. We show that the explanations are consistent and that the accuracy of SVM models are bounded by the accuracy of explanation components. In further experiments, generated explanation components are used to classify the sample data. The results show that the generated explanations display a high level of fidelity.

Index Terms—Text summarisation; Rule Extraction; Data Mining; Support Vector Machines; Explanation;

I. INTRODUCTION

A. Motivation

In recent years, machine learning techniques such as support vector machines (SVMs) have shown significant potential to the practice of medicine and to the psychiatric classification [1]. The application of machine learning techniques in psychiatric diagnosis has significant merits because of the lack of standardised biological diagnostic tests. Conventionally, expert psychiatrists, consciously and unconsciously analyse the language of their patients to make a clinical diagnosis using diagnostic classification schemas, such as DSM IV [2] and ICD 10 [3]. To provide an objective clinical diagnosis, SVMs have been applied to conversations recorded from patients and clinicians [1].

However, an explanation capability is crucial in security-sensitive domains such as medical applications. Although support vector machines (SVMs) have shown superior performance in a range of classification and regression tasks, SVMs like artificial neural networks (ANNs) lack an explanatory capability. There is a significant literature on obtaining human-comprehensible rules from SVMs and ANNs in order to explain “how” a decision was made or “why” a certain result was achieved [4], and this paper proposes a novel approach for SVM classifiers. The experiments describe a first attempt to generating text summaries for explaining “why” certain text contents are depressing or funny using a decompositional approach. Learned model parameters are analysed to select informative feature components and filtering is applied to select subsets of more relevant and reliable components.

To generate textual explanations, each text sample is converted into a set of basic concept-constructs called predicates using a natural language parser. The generated explanation components are then used to score relevant predicates using a similarity measure function, which is based on a common sense database called ConceptNet. The ranked predicates are then used to generate textual explanations of SVM classifications. Unlike previous text summarisation approaches, the generated text summaries explain why the particular sample is classified as positive or negative.

Furthermore, we show that generated explanations (sets of selected informative feature components) are consistent in the sense that an explanation term does not appear in two separate explanations for explaining two opposite class samples. We define “accuracy of the explanation terms” and show that the accuracy of an SVM model is bounded by accuracy of explanation terms. That is, the accuracy of an explanation term is always greater or equal to the accuracy of an SVM model.

B. Background

This work is based on text mining and in particular text classification. The following section provides a brief overview of the core techniques, focussing on support vector machines (SVMs), the significance of generating human-comprehensible explanations from SVMs, and what it means to explain the decision-making process of a machine learning system to a human user who may not be an expert on artificial intelligence.

1) Support Vector Machines: Cortes & Vapnik [5] introduce support vector machines which are a novel approach to machine learning. SVMs are based on the structural risk minimization principle in order to overcome the overfitting problems. Support vector machines find the hypotheses out of the hypothesis space $H$ of a learning system which approximately minimises the bound on the actual error by controlling the empirical error using training samples and the complexity
of the model using the VC-dimension of $H$. SVMs are very universal learning systems [6]. In their basic form, SVMs learn maximal margin hyperplanes (linear threshold functions). A hyperplane can be defined by a weight vector $w$ and a bias $b$:

$$w \cdot x + b = 0$$

The corresponding threshold function for an input vector $x$ is then given by:

$$f(x) = \text{sign}(w \cdot x + b)$$

However, it is possible learn polynomial classifiers, radial basis function (RBF) networks and three or more layered neural networks by mapping input data $x$ to some other (possibly infinite dimensional) feature space $\phi(x)$ and using kernel functions $K(x_i, x_j)$ to obtain dot products, $\phi(x_i) \cdot \phi(x_j)$, of feature data.

2) Explanations - The foundation: To illustrate why it is significant to provide explanation capability to SVMs, let us consider the case where a medical doctor tells a patient of a diagnosis that whether the result is positive or not. It is a social norm that doctors usually also include explanations of the diagnosis result. For instance, the explanations may be a deductive argument which comprises a list of observed symptoms of the patient, a list of possible causes, and modus ponens (the rule of inference) for deriving the conclusion.

Thagard and Litt [7] illustrate several major approaches to generating explanations. The classical view is that explanation is a deductive argument including background knowledge and inference rules such as modus ponens. The inference rules allow the sequential application of “if-then-else” statements in order to justify an explanatory target. Whenever no precise knowledge is available, explanatory schemas or probabilistic rules can be used.

Cawsey [8] is using a very simple definition of explanation: “In general an explanation is something which makes some piece of knowledge clear to the hearer. … The explanation is complete when the hearer is satisfied with the reply and understands the piece of knowledge” [8]. Hence, explanation is based on an “information need”.

3) Generating Explanations from SVMs: Much of works trying to provide an explanation capability to SVMs have been on rule extraction [4] following the footsteps of the earlier effort to obtain human-comprehensible rules from artificial neural networks (ANNs). One approach of classifying rule-extraction is the translucency dimension of classification, which comprises of the decompositional and pedagogical (or learning based) techniques [9].

The decompositional approach relies on the degree to which the internal representation of the ANN is accessible to the rule extraction technique. The basic strategy of decompositional techniques is to extract rules at the level of each individual hidden and output unit within the trained ANN. In general, decompositional rule extraction techniques incorporate some form of analysis of the weight vector and associated bias (threshold) of each unit in the trained ANN. Then, by treating each unit in the ANN as an isolated entity, decompositional techniques initially generate rules in which the antecedents and consequents are expressed in terms which are local to the unit from which they are derived.

In contrast to the decompositional approaches, the strategy of the pedagogical approaches is to view the trained ANN at the minimum possible level of granularity i.e. as a single entity or alternatively as a “black box”. The focus is on finding rules that map the ANN inputs (i.e. the attribute/value pairs from the problem domain) directly to outputs [10]. In addition to these two main categories, Andrew et al. [9] also proposed a third category which they labelled as “eclectic” to accommodate those rule extraction techniques which incorporate elements of both the decompositional and pedagogical approaches.

4) Translucency and rule quality applied to rule extraction from SVMs: It is very easy to illustrate the limitations of current studies on rule extraction from SVMs by use of an example: text classification. SVMs can achieve good performance with very simple text representation formats such as the “bag-of-words” (BOW) technique. BOW uses a document-term matrix such that rows are indexed by the documents and columns by the terms (e.g. words). SVMs allow the classification of texts of differing lengths; hence, document vectors may differ greatly in the number of elements.

A disadvantage of the BOW representation is that after successful classification, it may not be obvious what has been learned. For instance, an author may have a preference for certain topics and as a result, an SVM trained on an authorship identification problem in reality may perform topic detection. This problem has lead to various techniques to eliminate content from the BOW input, for instance by replacing content words with lexical tags (categories).

Given the fact that it is not at all obvious what contributes to classification in case of a BOW input representation, rule extraction from support vector machines is presented with a special opportunity. However, the number of features can be very large: e.g. all words that exist in a given natural language. While a combination of words constitutes meaning in a natural language, BOW and hence classification is based on words in isolation. This is a significant problem with regard to rule quality: The antecedents in a rule include individual words completely out of context. As the set of antecedents includes completely unrelated words, human or semantic comprehensibility is low.

5) Evaluation of the quality of extracted rules: Rule extraction from neural networks adopted criteria for the quality of the extracted rules. The set of criteria for evaluating rule quality includes (Andrew et al. [9]):

1. accuracy
2. fidelity
3. consistency, and
4. comprehensibility of the extracted rules.

A rule set is considered to be accurate if it can correctly classify a set of previously unseen examples from the problem domain [10]. Similarly a rule set is considered to display a high level of fidelity if it can mimic the behavior of neural network from which it was extracted by capturing all of the
information represented in the ANN. An extracted rule set is deemed to be consistent if, under differing training sessions, the neural network generates rule sets which produce the same classifications of unseen examples. Finally the comprehensibility of a rule set is determined by measuring the size of the rule set (in terms of the number of rules) and the number of antecedents per rule [10].

C. Overview

The reminder of this paper summarises experiments and their results: text classification, explanation generation for classification results, and technical details of methods with statistical analysis on the model parameters that are generated for depression poems. Then, in Section IV, we show how explanation terms can be used to generate textual summary of the classification results.

II. EXPERIMENTAL EVALUATION

A. Methodology - Explanation Term Generation

A preliminary study has been undertaken on generating explanations of depression poems and classifications of online text messages. The depression poems were obtained from the Internet (PoetryAmerica.com) and comprise of a total of 76 poems: 56 depression poems and 20 funny poems. The online text messages were obtained from Usenet news groups and comprise of a total of 350 sentences: 297 open questions and 53 closed questions. The resulting text documents are represented as attribute-value vectors (“bag of words” representation) where each distinct word corresponds to a feature component whose value is the frequency of the word in the text sample. Values were transformed with regard to the length of the sample. For the poem data set, functional words were removed and each word is converted into its lemma form (its base form without inflections). In addition, words that were not present in ConceptNet\(^1\) were removed. For the question data set, all words were used. In summary, input vectors for machine learning consist of attributes (the words used in the sample) and values (the transformed frequency of the words). Outputs are depression versus funny and open question versus closed question, that is, binary decision tasks were learned. Clearly, the expressive power of the resulting explanations is limited by this “bag of words” representation.

For LOO (leave-one-out) cross validation, 76 and 350 SVM models were generated using the linear kernel for the poem text data set. SVM models: the accuracy of each explanation terms increases (decreases) as the accuracy of the SVM models increases (decreases). SVM models: the accuracy of each explanation terms increases (decreases).

1) Explanation A comprising of all the feature components contributing to the decision value \(d(x)\).

2) Explanation B comprising of top-contributing feature components that are sufficient to classify the features.

3) Explanation C comprising of top-contributing feature components that also have their sensitivity values, \[|\frac{\partial d(x)}{\partial x_j}|\text{, greater than a set threshold value }\tau\text{.}

Technical details on generating each explanation types are described in Section III. This approach is clearly a decompositional approach: analysis on the model parameters to select informative components and selecting subsets of more relevant components. Figure 1 summarises the significance of each types of explanations. It plots sensitivity, contribution, and word rank of all feature components of the depression-poem text data set. It shows that sample feature components having higher ranking order (more frequent words) and higher sensitivity values tend to have larger contribution values. This suggests that feature components having higher sensitivity values and higher ranking order provide greater information in decision making than other feature components. It also shows that most of large contributions are made by more frequent words (high rank words).

In Section III, we show that the accuracy of each explanation terms is positively correlated to the accuracy of the SVM models: the accuracy of each explanation terms increases (decreases) as the accuracy of the SVM models increases (decreases).

B. Result of Explanation Generation

Support vector machines trained on the poem and online text message data sets achieved accuracy of 94% and 98%, respectively. The sensitivity values are adjusted manually to obtain reasonable amount of explanation terms for Explanation type C. Sample explanations of a depressive poem are provided below:

\[d(x) = w \cdot \phi(x) + b = \sum_{i \in SV} \alpha_i y_i K(x_i, x) + b\]

where \(x_i\) are support vectors and \(x\) is the feature vector, \(\alpha_i\) are Lagrangian multipliers, and \(b\) is the offset. The support vectors and the Lagrangian multipliers can be found by solving a quadratic programming problem. A popular setup of an SVM quadratic programming problem that allows some classification errors in the solution [5] is shown below:

\[
\min \left( \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \right)
\]

subject to: \(y_i (w^T \phi(x_i) + b) \geq 1 - \xi_i\), \(\xi_i \geq 0\)

where \(C\) is the penalty to errors and \(\xi_i\) are slack variables for allowing errors. This particular formula isn’t important here. The important thing is that the antecedent of the rule of inference is \(d(x) \geq 0\). That is, if \(d(x) \geq 0\), then the feature vector \(x\) is positive or else negative. Unlike previous rule extraction approaches, for each sample \(x\), we formulate textual summaries to explain why \(d(x) \geq 0\) or \(d(x) < 0\).

The first step in generating textual explanations for decisions is extracting feature components that contribute to the decision value \(d(x)\). We define three types of explanation sets comprising of a subset of feature components \(x_j\) of each feature vector \(x\):

1) Explanation A comprising of all the feature components contributing to the decision value \(d(x)\).

2) Explanation B comprising of top-contributing feature components that are sufficient to classify the features.

3) Explanation C comprising of top-contributing feature components that also have their sensitivity values, \[|\frac{\partial d(x)}{\partial x_j}|\text{, greater than a set threshold value }\tau\text{.}

\(^1\)Used ConceptNet v2.1 from the Common Sense Computing Initiative at the MIT Media Lab (http://csc.media.mit.edu).
The numbers \((d, q)\) in the brackets indicate relative contribution values \(d\) to the decision value \(d(x)\) and sensitivity values \(q\), respectively. For positive cases, if the sensitivity value of a feature component is positive, it means that the increase in the frequency of the feature component contributes to the decision value. Sensitivity-filtering (Explanation C) eliminates some of less sensitive terms (bold-faced terms) from Explanation B.

Sample explanations of a funny poem are provided below:

1. Explanation A: dont (56 53), call (23 44), tear (21 50),
know (17 18), fall (16 37), leave (16 35), cut (16 27),
dark (14 24), sad (10 17), cold (10 16), face (10 20),
smile (8 13), belong (5 7), letter (4 7), star (4 7), say (4 9),
grav(e (2 3), shell (2 3), mold (1 2), useless (1 1)
2. Explanation B: dont (56 53), call (23 44), tear (21 50),
know (17 18), fall (16 37), leave (16 35), cut (16 27),
dark (14 24), sad (10 17), cold (10 16), face (10 20)
3. Explanation C: dont (56 53), call (23 44), tear (21 50),
fall (16 37), leave (16 35), cut (16 27)

The numbers \((d, q)\) in the brackets indicate relative contribution values \(d\) to the decision value \(d(x)\) and sensitivity values \(q\), respectively. For positive cases, if the sensitivity value of a feature component is positive, it means that the increase in the frequency of the feature component contributes to the decision value. Sensitivity-filtering (Explanation C) eliminates some of less sensitive terms (bold-faced terms) from Explanation B.
For linear SVMs, we can easily obtain the contributions of the $j$-th feature component.

**Lemma 2:** Let $D(x) = \phi(x) - C$ be the deviation of a feature vector from a centroid $C$ of the population which is on a linear SVM hyperplane: $w \cdot x + b = 0$. Then, the contribution $D(x)_j$ of the $j$-th feature component $x_j$ of a feature vector $x$ to the deviation is proportional to the contribution $d(x)_j$ of the $j$-th component to the decision value $d(x) = w \cdot x + b$:

$$d(x)_j = a_j D(x)_j$$

where $a_j$ is a constant.

**Proof:** According to Lemma 1, the decision value of the feature vector is proportional to the projection of the deviation to a normal vector of the hyperplane:

$$d(x) = w \cdot D(x)$$

If $K$ is the linear kernel, we can estimate the contribution of each $j$-th feature component $x_j$ as follows:

$$d(x)_j = \sum_{i \in SV} \alpha_i y_i x_{i,j} (x_j - C_j)$$

$$= (x_j - C_j) \sum_{i \in SV} \alpha_i y_i x_{i,j}$$

$$= w_j D(x)_j$$

where $w_j = \sum_{i \in SV} \alpha_i y_i x_{i,j}$ is the $j$-th component of the weight vector $w$.

For linear SVMs, we can use Lemma 2 to generate explanations.

### A. Consistency of Explanations

By the definition of Explanation A, if a sample is classified as positive (negative), a feature component $x_j$ is included in Explanation A as an explanation term if $d(x)_j > 0$ ($d(x)_j < 0$), respectively. If a feature component $x_j$ is included in an explanation with $D(x)_j > 0$ for a sample, it means that the feature component is included as an explanation term because the feature component appears more frequently in the sample than the centroid $C_j$ of the feature component. Naturally then, for our explanations to be consistent, the same feature component should not be included in an explanation to explain an opposite class. We now show that Explanation A, B, and C are consistent. Consistency is one of the criteria for evaluating rule quality includes (Andrew et al. [9]).

**Theorem 1:** Explanation A, B, and C are consistent. For explanation A, B, and C, the following holds: (1) If a feature component $x_j$ appear in an explanation for a sample classified as positive with $D(x)_j > 0$ (i.e., increase in the frequency of feature component $x_j$), it does not appear in an explanation for any samples classified as negative with $D(x)_j < 0$. (2) if a feature component $x_j$ appear in an explanation for a sample classified as negative with $D(x)_j > 0$, it does not appear in an explanation for any samples classified as positive with $D(x)_j > 0$. (3) if a feature component $x_j$ appear in an explanation for a sample classified as positive with $D(x)_j < 0$ (i.e., decrease in the frequency of feature component $x_j$), it does not appear in an explanation for any samples classified as negative with $D(x)_j < 0$. (4) if a feature component $x_j$ appear in an explanation for a sample classified as negative with $D(x)_j < 0$, it does not appear in an explanation for any samples classified as positive with $D(x)_j < 0$.

**Proof:** We show that condition (1) holds. By condition (1), $x_j$ is a sample classified as positive and included in Explanation A with $D(x)_j > 0$. By the definition of Explanation A, $d(x)_j > 0$ and $w_j > 0$. Suppose $x_j$ is included in Explanation A with $D(x)_j > 0$, then by definition of Explanation A, $d(x)_j < 0$. However, we have $d(x)_j = w_j D(x)_j > 0$, contradiction! The rest of the conditions can be proved similarly.

### B. Accuracy of Explanation Terms

Another important criteria for evaluating rule quality (Andrew et al. [9]) is accuracy. Conventionally the accuracy of a binary classifier is defined as follows:

$$A_M = \frac{TP + TN}{N}$$

where $N$ is the total number of samples, $TP$ is the number of true-positive classification results and $TN$ is the number of true-negative classification results. Unfortunately, it is not that straightforward to define “accuracy of explanation terms” and there can be many different definitions depending on one’s definition on what it means by an explanation is accurate. However, we find that the following definition is the most natural way of defining the concept of “accuracy of explanation terms”. We start by defining “the error rate of an explanation term”.

**Definition 1:** The error rate of an explanation term $x_j$ is the number of times that $x_j$ is used incorrectly in an explanation divided by the number of explanations generated.

For Explanation A, B, and C, the number of times that $x_j$ is used incorrectly in an explanation is the sum of the number of times that $x_j$ is used for explaining negative samples with $d(x)_j > 0$ and the number of times that $x_j$ is used for explaining positive samples with $d(x)_j < 0$.

**Definition 2:** Let $M$ be a linear SVM model. Then, the empirical error rate $E_{j,M}$ of an explanation term $x_j$ for Explanation A of the SVM model is

$$E_{j,M} = \frac{FP_{d(x)_j > 0} + FN_{d(x)_j < 0}}{N}$$

where $N$ is the total number of samples, $FP_{d(x)_j > 0}$ is the number of explanations containing $x_j$ for false-positive classification results and $FN_{d(x)_j < 0}$ is the number of explanations containing $x_j$ for false-negative classification results. The accuracy of an explanation term $x_j$ is simply $1 - E_{j,M}$:

$$A_{j,M} = 1 - E_{j,M}$$

With these definitions we can now show that the accuracy of an SVM model is bounded by the accuracy of explanation terms.
Theorem 2: Let $M$ be a linear SVM model. Then, the accuracy $A_M$ of the SVM model $M$ is bounded by the accuracy $A_{j,M}$ of explanation terms $x_j$:

$$A_M \leq A_{j,M}$$

Proof: The accuracy of an explanation term $x_j$ is defined as follows

$$A_{j,M} = 1 - E_{j,M} = 1 - \frac{FP_{d(x_j)>0} + FN_{d(x_j)<0}}{N}.$$

By definition, the accuracy of the SVM model $M$ is:

$$A_M = \frac{TP + TN}{N} = 1 - \frac{FP + FN}{N}$$

where $FP$ is the number of false-positive classification results and $FN$ is the number of false-negative classification results. By definition, the set of elements included in $FP_{d(x_j)>0}$ ($FN_{d(x_j)<0}$) is a subset of $FP$ ($FN$), respectively. Thus, $FP_{d(x_j)>0} \leq FP$ and $FN_{d(x_j)<0} \leq FN$. Thus, the following holds for any explanation term $x_j$:

$$\frac{FP_{d(x_j)>0} + FN_{d(x_j)<0}}{N} \leq \frac{FP + FN}{N}$$

Therefore, $A_M \leq A_{j,M}$ for any explanation term $x_j$. Figure 2 illustrates the proof for two dimensional feature space. In the figure, the samples are drawn from a mixture of two multivariate Gaussian distribution: $G+$ and $G−$. It illustrates that $FP_{d(x_j)>0}$ (the shaded area B) is smaller than $FP$ (the sum of two areas A and B).

For our experiments, we calculate the centroid components $C_{sv,j}$ of the support vectors as estimated population-centroid components:

$$C_{sv,j} = \frac{1}{N_{sv}} \sum_{i \in SV} x_{i,j}$$

Now, for a feature vector $x$, we can explain why a sample is positive (negative) by listing the feature components that contribute to the decision value. That is, we can rank the feature components of a feature vector according to the amount of contributions made by the feature components. This is used as the basis of the explanation type A. We can also calculate the sum of all negative (positive) contributions and choose the top $N$ positive (negative) contributions that are sufficient to push the decision value to positive (negative). This is used as the basis of the explanation type B.

1) Experimental Results of Accuracy of Explanation Terms: Figure 3 shows the distribution of the explanation-term accuracy of the poem text data. As predicted by Theorem 2, the minimum accuracy value is 0.94 which is the accuracy of the corresponding SVM model.

C. Fidelity of Explanations

In order to test the explanation capability of the explanation terms, we used all the explanation terms included Explanation A to generate new feature set for the poem text data. By using the explanation terms for generating the new feature set, the vocabulary size is reduced from 1410 to 554. Support vector machines trained on the new feature set achieved accuracy of 87% and AUC (Area Under the Curve) of 0.89. The corresponding ROC curve is shown in Figure 4. The ROC curve for the SVM model using the full vocabulary is shown in Figure 5. This clearly shows that the explanation terms used in the explanations have good explanation capability. That is, the explanations display a high level of fidelity.
Fig. 4. True-positive rate vs. false-positive rate of a linear support vector machine using an explanation-term vocabulary.

Fig. 5. ROC curve of the SVM model of the depression-poem data. This illustrates that performance indicators can be adjusted by moving the threshold of the decision value. Specificity (recall) rate is increased from 0.549 to 0.85 by moving the decision threshold from 0 to 0.162. This effectively moves the estimated centroid of the population to produce more accurate explanations for imbalanced data.

because the explanation-term set can mimic the behavior of the SVM model from which the explanation terms are extracted. According to Theorem 1, the explanation terms are also consistent. That is, if an explanation term is used to explain a positive case, then the same explanation term is not used to explain a negative case.

D. Optimisation for Imbalanced Data

SVMs have been successfully applied to many text classification tasks, for example to determine mental health problems using transcribed speech samples [1]. However, very few data sets have the equal number of samples: often the number of positive and negative samples are very different. This is particularly true for medical data, where we have either very few positive samples because, for example, positive cases are very rare or very few negative samples. For example, Autism assessment records contains very few negative cases because most of the patients are already highly suspected of the symptoms when they are referred to Autism specialists. This imbalance in data sets can greatly affect on the performance of machine learning algorithms. This imbalance in data can affect the accuracy of the estimated population-centroid, and thus explanations can become unreliable.

Various adjustments to SVMs are proposed to improve the performance of SVMs with imbalanced data [11], [12]. Most of the proposed approaches are based on the idea that the locations of the SVM hyperplanes can be adjusted to account for imbalanced data. One approach is to use separate cost factor measures $C_+$ and $C_-$ for positive and negative samples, respectively [11]. Another approach is to adjust the bias term $b$ [12] after n-fold cross validation to find optimal performance indicators.

In our experiments, we use the approach of adjusting the bias term $b$ after n-fold cross validation. We use receiver-operating-curve (ROC) and balanced accuracy (BAC) as the heuristics for finding the optimal adjustment amount of the bias term.

$$BAC = \frac{specificity + sensitivity}{2}$$

$$specificity = \frac{true\_negative\_rate}{P(O_-|L_-)}$$

$$sensitivity = \frac{true\_positive\_rate}{P(O_+|L_+)}$$

Figure 5 shows the ROC curve of the LOO (leave-one-out) cross validation results of the depression-poem data. Using the default bias, the model had specificity value of 0.59. By adjusting the bias to $b = 0.17628$ (i.e., a sample is classified as positive if $d(x) > 0.17628$), we obtained the specificity and sensitivity values of 0.8 and 0.96, respectively.

Adjusting the bias term moves the hyperplane along the weight vector $w$. This movement has to be considered in calculating the deviation.

$$D'(x) = \phi(x) - \left(C_{sv} + \Gamma\right)$$

where $\Gamma$ is the adjustment to the centroid. If the adjustment value to the bias term is $\delta$ (i.e., a sample is classified as positive if $d(x) > \delta$), $\Gamma$ is defined as follows

$$\Gamma = \delta \frac{w}{||w||^2} = \delta_n w$$

where $\delta_n = \delta / ||w||^2$ is the normalised adjustment of the hyperplane. Then, we can obtain the new decision value $d'(x)$ using the adjusted deviation $D'(x)$:

$$d'(x) = w \cdot (\phi(x) + b - \delta)$$

$$= w \cdot (\phi(x) - (w \cdot (D(x) + C_{sv} - \Gamma) + b)$$

$$\approx w \cdot (D(x) - \Gamma)$$

$$= w \cdot (\phi(x) - (C_{sv} + \Gamma))$$
Fig. 6. Distribution of sensitivity values of the poem text data set.

\[ D'(x) = w \cdot [\phi(x) - C_{sv} - \delta_n w] \]

If \( K \) is the linear kernel, we can estimate the contribution of each \( j \)-th feature component \( x_j \) as follows:

\[ d'(x)_j = \sum_{i \in SV} \alpha_i y_i x_{i,j} (x_j - C_{sv,j} - \delta_n x_{i,j}) \]

E. Filtering Explanations with Sensitivity

Training a support vector machine for a data set of interest generates a hyperplane, which can be used to obtain the distance of a feature vector to the hyperplane to classify. The distance is normal to the hyperplane and thus the importance of a feature component can be measured as the rate of change of the distance with respect to the feature component. This can be easily obtained for linear classifier as follows:

\[ \frac{\partial d(x)}{\partial x_j} = \sum_{i \in SV} \alpha_i y_i x_{i,j} = w_j \]

where \( d(x) \) is the distance of feature \( x \) to the hyperplane, \( x_j \) is the \( j \)-th component of the feature \( x \), \( x_{i,j} \) is the \( j \)-th component of a support vector \( x_i \), and \( w_j \) is the \( j \)-th component of the weight vector \( w \). As we can see from the above equation, the importance of the \( j \)-th component for the hyperplane is the sum of \( j \)-th component of the support vectors multiplied by the class label and the Lagrange multipliers.

Conjecture 1: As the sensitivity of the \( j \)-th feature component increases (decreases), the error rate of the \( j \)-th explanation \( (1 - \text{accuracy}_j) \) decreases (increases).

Figure 2 illustrates that as the sensitivity of the \( j \)-th feature component increases (decreases), the error rate of the \( j \)-th explanation \( (1 - \text{accuracy}_j) \) decreases (increases).

Figure 6 shows a histogram of sensitivity values of feature elements for the poem text data set. Greater population is centred at sensitivity value 0. This suggests that feature components having higher sensitivity values will provide more information on the decisions. This is also suggested in Figure 1 that most of the contributions are made by terms with higher sensitivity values.

Figure 7 shows the relationship between word ranks and sensitivity values of sample feature components. Feature components with higher sensitivity values tend to have higher ranking order (i.e., lower rank value). The relationship between the word rank and sensitivity for the poem text data set can be summarised as follows:

\[ \text{rank} \leq \alpha \frac{1}{|\text{sensitivity}|} \]

for some positive constant \( \alpha > 0 \). A similar relationship is also observed between the word rank and contribution across different text data sets (\( \beta > 0 \)):

\[ \text{rank} \leq \beta \frac{1}{|\text{contribution}|} \]

Whereas the contribution and sensitivity are proportional

\[ |\text{contribution}| \leq \gamma |\text{sensitivity}| \]

which strongly suggests the two-step filtering (Explanation type B and Explanation type C).

For non-linear cases, we have to obtain partial derivatives of kernels. As an example, let us consider the polynomial kernel:
The feature component $j$ and the explanation set $E$ of the term $k$ of the predicates. We then calculate scores of predicates. The score with their corresponding antecedents to improve readability of identified by using ConceptNet and the pronouns are replaced is broken into a set of predicates (verb-subject-object tuples) instructors of text. Instead of using sentences, each sentence where $k$ contributions made by parts of the sentences. The score of the in this approach, each sentence is given scores by determining the text summary is for explaining why the particular sample classification result. Unlike previous text summarisation approaches, the textual information to measure the relevancy of each parts of the text is classified as positive or negative.

The second method we developed uses more basic con-

• $u_j$ ("end") = 0.269967456035
• $u_j$ ("miss") = 0.75678954875
• $u_j$ ("cry") = 0.599278222108
• $u_j$ ("tear") = 0.168901775853

The second method we developed uses more basic constructors of text. Instead of using sentences, each sentence is broken into a set of predicates (verb-subject-object tuples) that make up the sentences. Antecedents of pronouns are identified by using ConceptNet and the pronouns are replaced with their corresponding antecedents to improve readability of predicates. We then calculate scores of predicates. The score of the $k$-th predicate in a sample is defined as follows:

$$ s_k = \frac{1}{|S_k|} \sum_{j \in E_k} \sum_{t \in S_k} u_j(t) d(x)_j q(x)_j $$

where $u_j(t)$ is the utility function that measures how close the term $t$ in the sentence $S_k$ is to feature component $j$ in an explanation set $E$, $d(x)_j$ is the amount of contribution of the feature component $j$, and $q(x)_j$ is the sensitivity of the feature component $j$. The text summary is then generated by selecting a subset of sentences from the text using the scores as relevance measures. ConceptNet analogy space is used as the utility function. The following is example similarity values for $j$ = "frustration":

- $u_j$ ("end") = 0.269967456035
- $u_j$ ("miss") = 0.75678954875
- $u_j$ ("cry") = 0.599278222108
- $u_j$ ("tear") = 0.168901775853

This shows that the importance of the $j$-th input feature component for the hyperplane is a combination of other input feature components weighted by support vector components. In this case, one way of obtaining sensitivity values would be averaging the sensitivity values over a range of selected points, such as support vectors.

IV. CONTEXTUAL TEXT SUMMARISATION: APPLICATION OF EXPLANATION GENERATION

The explanations generated provide relevancy of each feature components to the particular classes. We can use this information to measure the relevancy of each parts of the text data to generate a textual summary with regard to a classification result. Unlike previous text summarisation approaches, the text summary is for explaining why the particular sample is classified as positive or negative.

We start with a simple approach of generating a textual summary. The first method is scoring each sentences in a sample using the explanation terms generated for the sample. In this approach, each sentence is given scores by determining contributions made by parts of the sentences. The score of the $k$-th sentence in a sample text is defined as follows:

$$ s_k = \frac{1}{|S_k|} \sum_{j \in E_k} \sum_{t \in S_k} u_j(t) d(x)_j q(x)_j $$

where $u_j(t)$ is the utility function that measures how close the term $t$ in the sentence $S_k$ is to feature component $j$ in an explanation set $E$. Similarly to sentence based summarisation, the text summary is then generated by selecting a subset of the predicates using the scores as relevance measures.

A. RESULT OF TEXT SUMMARISATION

The following is a sample sentence-based text summary generated from a depression poem comprising of 28 sentences. The numbers are score values given to each sentence:

1) 0.187 This thought comes almost everyday
2) 0.166 We have become close friends as we were one in the same
3) 0.172 But you have never been able to see
4) 0.194 You may ask and look concerned wanting to know why I cry
5) 0.164 But do you really want to that I wish for me to die
6) 0.178 Can you handle the pain I have felt and dealt with by myself for many years
7) 0.166 I cant see the joy I once felt
8) 0.164 I just turn my head as quick as possible for I do not want see what Im actually trying to hide
9) 0.164 In the end will you miss me after I have cried my last tears

The following is a sample predicate-based text summary generated for the same depression poem. The numbers are score values given to each predicates:

1) 0.198 come everyday
2) 0.228 long someone
3) 0.222 they need me
4) 0.228 long i
5) 0.224 i protect me
6) 0.228 you be able
7) 0.222 you ask
8) 0.224 i feel
9) 0.228 you have never been able to see
10) 0.224 i deal myself year
11) 0.194 i deal myself year
12) 0.250 i feel

V. DISCUSSIONS AND FUTURE WORK

This is the first report of a novel approach to generating high-quality textual explanation of psychiatric classification, in particular explaining depression poem text contents. We have shown that the explanations are consistent, accurate, and display a high level of fidelity. Furthermore, we have shown that the approach can be applied to imbalanced data by adjusting SVM hyperplanes and centroids using ROC curves. To improve the comprehensibility of explanations, the input text is parsed into predicates and scored using a common sense database called ConceptNet. We believe that this approach overcomes the subjectiveness of measuring comprehensibility. Considerable further research is required. The approach of extracting some piece of knowledge using machine learning in explaining psychiatric assessments has
the potential to improve the usability of machine learning techniques in the medical and security domains. This approach of extracting explanations using some form of analysis on machine learning and associated parameters can be further expanded by using alternative feature representations of text data sets, such as concept terms or semantic terms. In other, as yet unpublished work, the authors have already used these approaches to generate explanations with reasonable success. There is massive potential of incorporating these sophisticated information extraction technologies within psychiatry and in medicine more generally.

REFERENCES