2021 International Conference on Data Mining Workshops (ICDMW)

Sentiment Analysis Using Part-of-Speech-Based Feature Extraction and Game-Theoretic Rough Sets

Yixing Chen and JingTao Yao
Department of Computer Science
University of Regina
Regina, Canada
[ycf667, jingtao.yao]@uregina.ca

***Abstract*—Sentiment analysis, one of the most trending natural language processing tasks, is used to mine opinions or sentiments from a given text. Two significant challenges of sentiment analysis are 1) complexity in data pre-processing caused by the high dimensionality of textual data; 2) uncertainty in classifying sentiment polarities due to the ambiguity of natural languages. To address these issues, we propose a model using part-of-speech-based feature extraction to reduce dimensionality and game-theoretic rough sets (GTRS) to establish a balance between the accuracy and coverage trade-off. We evaluate this model with three different sizes of datasets (Yelp reviews, IMDB movie reviews, and Amazon product reviews). The experiment results show that the proposed model outperforms Pawlak's rough set model and 0.5-probabilistic rough set model. In comparison with four traditional binary classification models (i.e., SVM, naïve Bayes, decision tree, and KNN), the proposed model also achieves higher accuracy rates. This research suggests that the proposed model is promising to deal with the complexity and uncertainty in sentiment analysis tasks.**

***Index Terms*—sentiment analysis, part-of-speech tagging, feature extraction, three-way classification, game-theoretic rough sets**

I. Introduction

Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions or sentiments expressed in written text [16]. Unlike traditional topic-based text classification that classifies texts based on their subject matter (e.g., sports, politics, or finance), sentiment analysis classifies texts into different sentiments (e.g., positive or negative sentiment), which can be expressed in a more subtle manner [19].

Sentiment analysis application domains include customer products, health care, tourism, hospitality, financial services, social events, and political elections. Sentiment analysis technique can provide opinions of others to businesses, organizations, government agencies, and even individual customers for better decision making [16].

There are two significant challenges in sentiment analysis tasks: complexity in data pre-processing and uncertainty in classifying sentiment polarities [11] [28]. One of the leading causes of the complexity is high dimensionality of textual data. The part-of-speech-based feature extraction that we propose is an intuitive and effective method to reduce the dimensionality in sentiment analysis tasks. This method transforms textual data into smaller feature sets based on words' part-of-speech category and converts each set of words to a corresponding sentiment polarity score. In this way, the high dimensionality is significantly reduced, and the information related to sentiment orientation is greatly preserved.

The uncertainty in classifying sentiment polarity is caused by the ambiguity of natural languages. Three-way classification has been studied and applied in various fields to address uncertainty in decision-making [25]. In sentiment analysis, by applying three-way classification, the instances are divided into positive, negative, and boundary regions, corresponding to three sentiment polarities: positive, negative, and neutral sentiments. In order to determine the threshold for dividing the three regions, we adopt game-theoretic rough sets (GTRS) [12]. With GTRS, we resolve the uncertainty by considering the trade-off nature between accuracy and coverage and finding the highest possible results of two criteria.

The proposed model is a hybrid approach of sentiment analysis that benefits both from a knowledge-based technique and a statistical technique [6]. In detail, the part-of-speech-based feature extraction leverages the knowledge bases exploited from linguistic corpora, and GTRS uses a statistical method for further classification. We experiment with three different sizes of datasets. The classification result is compared with Pawlak's rough sets, 0.5-probabilistic rough sets, as well as four traditional binary classification models: support vector machine (SVM), naïve Bayes, decision tree, and k-nearest neighbors (KNN). The result indicates that the proposed approach outperforms other methods with all three datasets.

II. Background Knowledge

This section explains the background knowledge of handling the complexity and uncertainty in sentiment analysis.

A. Part-of-Speech-Based Feature Extraction in Sentiment Analysis

Part-of-speech, often abbreviated as *POS* and also called word class, is a category of words that share certain grammatical properties. Words that are assigned to the same part-of-speech category generally perform similar functions in a sentence [15]. To avoid confusion with the abbreviation of positive region in the rough set community, we use "part-of-speech" throughout the paper.

In natural language processing tasks, the textual data need to be transformed into a numerical form (an attribute-value

2375-9259/21/$31.00 ©2021 IEEE
DOI 10.1109/ICDMW53433.2021.00020

representation) before being analyzed and computed by algorithms. In traditional text representation approaches, such as one-hot-encoding, TF-IDF (term frequency-inverse document frequency), and count vectorization, original datasets are vectorized and transformed into large and sparse datasets. The dimensionality (size of feature set) can be up to the size of the vocabulary across the entire dataset, generating huge computational cost [27].

Feature extraction is a general method of dimensionality reduction, which transforms input space into a lower dimension that preserves most of the relevant information [8]. In sentiment analysis tasks, extracting features for dimensionality reduction is to create a smaller set of features and feed this lower-dimensional data to classifiers to predict the sentiment polarity of the entire text. This smaller set of features contain the most sentiment-relevant words to represent the sentiment pattern of texts.

Part-of-speech-based feature extraction creates a smaller set of features based on the word's part-of-speech category. According to Cambridge Dictionary, nouns, verbs, adjectives, and adverbs are four major word classes in English. These four word classes can describe the behaviors or characteristics of an object, and they are likely to hold sentiment words that indicate sentiment polarities [13]. For instance, adjectives such as *nice* and *terrible*, adverbs like *happily* and *sadly*, verbs like *enjoy* and *hate*, nouns like *success* and *disaster* can explicitly point out the sentiment orientation in a text. Therefore, these word classes can be considered as the extracted features for sentiment analysis.

There are some sentiment analysis studies based on these four word classes. Among them, a large count of research extracts features based on the frequency or presence of words [18]. However, the frequency or presence-based solution processes all words in the corpus first, which generates a large scale of data for analysis. Therefore, this method requires a large processor memory to be implemented. Furthermore, solely predicting sentiment classes may not reveal sufficient sentiment-relevant information. To analyze finer-level details of the sentiment expressed, measuring the intensity of sentiment is of paramount importance [1].

In other research, words are extracted based on their sentiment intensity scores using SentiWordNet, such as the study [10]. However, this pioneer sentiment analysis tool is not a gold-standard resource and contains a noisy lexicon [14]. In our experiment, we adopt Valence Aware Dictionary for Sentiment Reasoning (VADER), which is a rule-based gold-standard lexicon [14]. It is trained on a human-validated sentiment lexicon and provides sentiment polarity (i.e., positive, negative, or neutral sentiment) and intensity scores. To calculate a text's sentiment score, VADER outputs a compound score by summing up all word's valence scores in their lexicon and normalizing them to between -1 and 1, corresponding to negative and positive polarities. Compared to SentiWordNet, VADER is faster and performs better in social media datasets [3]. Besides, VADER also outperforms other well-established sentiment analysis lexicons such as Linguistic Inquiry Word Count, General Inquirer, Affective Norms for English Words, SenticNet, Word-Sense Disambiguation using WordNet, and the Hu-Liu04 opinion lexicon [14].

Therefore, extracting four part-of-speech groups and converting them into sentiment scores using VADER as four features can reduce data dimensionality and preserve sentiment-related information in texts. As a result, we can save computational time without compromising the sentiment classification accuracy.

B. Three-way Classification and Probabilistic Rough Sets

In terms of the uncertainty in classifying sentiment polarities, several approaches are proposed to tackle this problem. The uncertainty theory using Delphi method is to invite domain experts to determine the uncertain instances in data [22]. A recent study adds an ambivalence category in sentiment classification and then further classifies based on transferring rules and human labeling [23].

There are also computational methods without human opinions involved, such as fuzzy theory and rough set theory [9] [28]. These methods analyze the data uncertainty from a mathematical perspective. They apply set theory and consider the uncertain data as a subset of the data. Then further explore the uncertain subset to make classification decisions.

Rough set theory, proposed by Pawlak [20], deals with the uncertain or incomplete information in an information table. The basic units in rough sets are equivalence classes where all instances with the same condition attributes are grouped as one equivalence class. Three-way classification in rough sets places the equivalence classes into three pair-wise disjoint regions: positive, boundary, and negative regions based on complete certainty. The drawback of Pawlak's rough sets is the zero-tolerance of uncertainty or error, as it discards all uncertain data. If data contain high uncertainty, the model can only classify a small fraction of data into positive or negative, leading to low data usability, also called data coverage.

Probabilistic rough set theory is proposed to solve the rigidness of Pawlak's problem [26]. This technique allows a certain degree of uncertainty in three-way classification. In detail, assume an instance x belonging to the universe U, is denoted as $x \in U$, and the equivalence class containing x is denoted as $[x]$. In probabilistic rough sets, a conditional probability $Pr(C|[x])$, is defined as the probability that an instance belongs to C (a set of criteria) given that this instance belongs to the equivalence class $[x]$.

Three-way classification has been applied in sentiment analysis tasks, such as [28], where the instances are divided into positive, negative, and boundary regions corresponding to three sentiment polarities: positive, negative, and neutral sentiment. In the proposed research, we define the concept "positive sentiment" as the criteria C, denoting $psent$ to distinguish with the notion "positive" in the concept "positive region". Hence, the conditional probability $Pr(psent|[x])$ is defined as the probability that an instance belongs to a positive sentiment given that this instance belongs to an equivalence class $[x]$, as shown in Equation (1), where $|\cdot|$ denotes the cardinality of a set.

$$Pr(psent|[x]) = \frac{|psent \cap [x]|}{|[x]|} \tag{1}$$

The pair of threshold (α,β), where $0 \leqslant \beta < \alpha \leqslant 1$, are to divide the positive, negative and boundary regions, denoting as POS, NEG and BND, respectively. Therefore, the three regions are a partition of the universe U, defined as:

$$\begin{aligned}\pi_{(\alpha,\beta)}(psent) = \{&POS_{(\alpha,\beta)}(psent),\\ &BND_{(\alpha,\beta)}(psent),\\ &NEG_{(\alpha,\beta)}(psent)\}\end{aligned} \tag{2}$$

Accordingly, the three-way sentiment classification is divided as:

$$\begin{aligned}POS_{(\alpha,\beta)}(psent) &= \{x|x \in U, Pr(psent|[x]) \geqslant \alpha\},\\ BND_{(\alpha,\beta)}(psent) &= \{x|x \in U, \beta < Pr(psent|[x]) < \alpha\},\\ NEG_{(\alpha,\beta)}(psent) &= \{x|x \in U, Pr(psent|[x]) \leqslant \beta\},\end{aligned} \tag{3}$$

As in Equation (3), for a given equivalence class $[x]$, if $Pr(psent|[x]) \geqslant \alpha$, then the instances in $[x]$ are accepted as positive sentiment; if $Pr(psent|[x]) \leqslant \beta$, then the instances in $[x]$ are rejected as positive sentiment, hence, negative sentiment; if $\beta < Pr(psent|[x]) < \alpha$, we defer to make decision due to the uncertainty or lack of information, so we consider them neutral sentiment.

C. Game-Theoretic Rough Sets

A challenging problem in the three-way classification based on probabilistic rough sets is finding the final threshold pair (α,β). GTRS provides a solution by applying game theory [12]. GTRS considers the trade-off between multiple criteria, and use game equilibrium to achieve final thresholds [30]. In detail, GTRS formulates games considering evaluation criteria as players, possible threshold pairs as strategies profiles. When the evaluation criteria in the game establish a game equilibrium, an optimal threshold pair is achieved. A GTRS game consists of three phases: game formulation, game implementation, and game result acquisition.

1) Game Formulation: There are three elements in a GTRS game: a set of players O, a set of strategies of each player S, the payoff functions of players u. Thus, for the game G, we have $G = \{O, S, u\}$. The three elements are defined as follows:

- Set of game players O: the evaluation criteria of three-way classification.
- Set of strategies S: the set of strategies or actions taken by each player. For example, each strategy is a unique threshold pair (α,β).
- Payoff functions u: payoff, or utility, is the result from a player performing a strategy, meaning the actual value of each evaluation criterion.

When the players (i.e., criteria) play the game, their strategies can impact their payoffs and form a trade-off. Their goals are to maximize their payoffs in the game.

2) Game Implementation: This phase requires observing and analyzing the impact of all eligible strategies on their corresponding payoffs.

3) Result Acquisition: The final phase is to analyze the game result and find the optimal threshold pair. Nash equilibrium is typically used to determine possible game outcomes in GTRS [30]. However, Nash equilibrium considers two players in a non-cooperative situation where players compete with each other, and each of them takes their own benefit as a priority. Therefore, Nash equilibrium can only achieve a balanced result as no one is willing to compromise. The proposed model uses Pareto efficiency as a game solution to find an optimal result by considering two players cooperating and jointly achieving their best possible outcomes. Our goal is to find a strategy profile where both players (i.e., accuracy and coverage) have payoffs as high as possible given their trade-off relation. Therefore, a complete observation of game strategies and corresponding payoffs is carried out based on a payoff table. The players can consider all payoffs and make an optimal choice.

GTRS has been applied to deal with information uncertainty such as Email spam filtering, satirical news detection, and recommender system [4] [29] [32].

III. Methodology

This section discusses in detail the main phases of the proposed model, including text cleaning, part-of-speech-based feature extraction, equivalence classes building, and sentiment analysis with GTRS.

A. Text Cleaning

We conduct several basic text cleaning procedures, such as removing URLs, removing punctuation, and tokenizing texts.

One important step during text cleaning is processing word contractions. Some words are in contracted forms, such as *don't*, *aren't*, and *wouldn't*, which are negation words. To avoid the loss of words' meaning during the text cleaning, we manually transform the contracted words into their base form before removing punctuation. For instance, the above contracted words are transformed to *do not*, *are not*, and *would not*.

Due to the case sensitivity of VADER, we skip the step of converting uppercase words to lowercase. In detail, VADER considers uppercase words to contain stronger and more intensive sentiment than lowercase words. For instance, based on VADER, the sentence *"It is great."* has a compound score of 0.6249, while *"It is GREAT."* has 0.7034, meaning the latter is more positive than the former. As such, we retain the original capitalization of words.

Regarding sentiment labels, we denote negative sentiment class as -1, positive as 1 for all experimental datasets.

B. Part-of-Speech-Based Feature Extraction

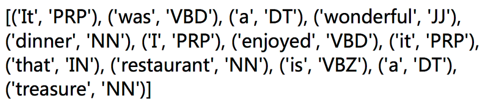
This phase has three steps: part-of-speech tagging, sentiment polarity scores converting, and data discretization.

TABLE I
EXTRACTED WORD CLASSES AND PART-OF-SPEECH TAGS

Word Class	Part-of-Speech Tag
Noun	NN, NNS, NNP, NNPS, *no*
Verb	VB, VBD, VBG, VBN, VBP, VBZ
Adjective	JJ, JJR, JJS
Adverb	RB, RBR, RBS, WRB

1) Part-of-Speech Tagging: We adopt the part-of-speech tagging tool in the Natural Language Toolkit (nltk) on the website *https://www.nltk.org/* to identify all the nouns, verbs, adjectives, and adverbs. The tags we select for the four word classes are shown in TABLE I. Each tag represents a form of the corresponding word class. For instance, "NN" and "NNS" represent singular and plural nouns, "NNP" and "NNPS" are singular and plural proper nouns; "VB" represents the base form of verbs, and the other tags starting with "VB" are other tenses of verbs; "JJ" represents adjective, and "JJR" and "JJS" represent comparative and superlative adjective; "RB", "RBR", and "RBS" correspond to adverb, comparative adverb and superlative adverb, "WRB" represents wh-adverbs.

One more case is the handling of *no*, because it is tagged as a determiner, not in any of the four word classes. We manually add this word into the group of noun as it is usually used combining with nouns.



```
[('It', 'PRP'), ('was', 'VBD'), ('a', 'DT'), ('wonderful', 'JJ'),
('dinner', 'NN'), ('I', 'PRP'), ('enjoyed', 'VBD'), ('it', 'PRP'),
('that', 'IN'), ('restaurant', 'NN'), ('is', 'VBZ'), ('a', 'DT'),
('treasure', 'NN')]
```

Fig. 1. An example of part-of-speech tagging

Taking the following sentence as an example: "*It was a wonderful dinner, I enjoyed it, that restaurant is a treasure*." After tokenized, part-of-speech tagger tags each word as shown in Fig. 1. Each text instance is converted into a list, where each list item is a tuple pair of a word and its corresponding tag.

2) Sentiment Polarity Scores Converting: As each word in a text instance is tagged, we search through the tags based on TABLE I. We then group four word classes and feed each group into VADER to obtain a compound score for each word class of a text instance. A score close to -1 indicates that the sentiment is close to negative, 1 indicates positive, and zero (0) is neutral.

TABLE II
AN EXAMPLE OF WORD CLASSES AND SENTIMENT SCORES

Word Class	Extracted Words	Sentiment Score
Noun	*dinner restaurant treasure*	0.296
Verb	*was enjoyed is*	0.5106
Adjective	*wonderful*	0.5719
Adverb	none	0

Taking the same example "*It was a wonderful dinner, I enjoyed it, that restaurant is a treasure*", four extracted word groups and their sentiment scores are shown in TABLE II. All words in this text instance that meet the extraction criteria are listed. Given that there are no adverbs in this sentence, VADER outputs the adverb sentiment score as zero. From this table, we can see that, except for the adverb group, the other word classes all achieve sentiment scores higher than zero, which indicates that the sentiment polarity of the entire text is very likely to be positive.

3) Data Discretization: All text instances are transformed to a four-dimensional representation, and consist of numbers ranging from -1 to 1. In order to be analyzed by rough sets, we apply an equal-width binning method to discretize the continuous values in each column. The width of each bin w is calculated as $w = (max - min)/5$, where max and min are the maximum and minim values in a column, and five is the number of bins that the values are divided. This binning method ensures that all bins have equal width. By ignoring the number of instances in each bin, we can preserve the original distribution of the sentiment scores.

TABLE III
A DEMO OF AN INFORMATION TABLE

Text	Noun	Verb	Adjective	Adverb	Label
Review_1	1	2	1	3	-1
Review_2	2	1	2	3	-1
Review_3	4	3	5	2	1
Review_4	5	4	2	2	1

As the extracted features are formulated, we combine them with sentiment labels, generating an information table for decision-making. TABLE III shows a demo of the information table (the dummy data values are for reference only). The first four columns are the extracted features, that is, condition attributes. The last column is the target attribute which is the sentiment label.

C. Equivalence Classes

In the datasets, we firstly build equivalence classes by finding the instances with the same condition attributes. Then in each equivalence class X_i, we calculate the equivalence class probability $Pr(X_i)$, and the conditional probability $Pr(psent|[x])$. The equivalence class probability $Pr(X_i)$ is the ratio of instances in the equivalence class X_i and the total number of all instances in the dataset. The conditional probability $Pr(psent|[x])$ is the proportion of instances that are positive sentiment in the equivalence class X_i, as in Equation (1).

D. Sentiment Analysis With GTRS

There are three steps in sentiment analysis with GTRS: game formulation, game implementation, and result acquisition, as follows:

1) Game Formulation: The game players, strategies, and payoff functions are defined as follows:

a) Game players: We use accuracy and coverage rates as game players, denoted as acc, and cov.

The accuracy rate is the ratio of the number of correctly classified instances and the number of all instances that can be classified, calculated as in Equation (4):

$$Acc_{(\alpha,\beta)}(psent) = \frac{|psent \cap POS_{(\alpha,\beta)}(psent)|+|psent^c \cap NEG_{(\alpha,\beta)}(psent)|}{|POS_{(\alpha,\beta)}(psent)|+|NEG_{(\alpha,\beta)}(psent)|} \tag{4}$$

Coverage rate is the ratio of the number of instances in positive and negative sentiment classes and the total number of instances in the dataset, calculated as:

$$Cov_{(\alpha,\beta)}(psent) = \frac{|POS_{(\alpha,\beta)}(psent)|+|NEG_{(\alpha,\beta)}(psent)|}{|U|} \tag{5}$$

b) Strategies: The set of strategies is $S = S_{acc} \times S_{cov}$, where S_{acc} and S_{cov} are the strategies taken by player acc and cov, respectively. In our experiment, the strategy takes different values of α and β. We set the initial threshold pair as (1,0), specifically, $\alpha = 1$ and $\beta = 0$. The change of two values is set to be 0.05 per step. In this GTRS game, we let player S_{acc} take the strategy of decreasing α, player S_{cov} takes the strategy of increasing β. We skip the pairs where α is less than β, meanwhile α should be no less than 0.5, and β no more than 0.5. Thus, we have strategies for each player as follows:

$$\begin{aligned} S_{acc} &= \{\alpha = 1, \alpha = 0.95, \alpha = 0.90, ..., \alpha = 0.50\}, \\ S_{cov} &= \{\beta = 0, \beta = 0.05, \beta = 0.10, ..., \beta = 0.50\} \end{aligned} \tag{6}$$

c) Payoff functions: The payoff of players is denoted as $u = (u_{acc}, u_{cov})$, where u_{acc} and u_{acc} are the payoff functions of player acc and cov. Assume a set of strategies has been decided, in other words, acc and cov both have chosen their α and β values, respectively. Then the payoffs of two players are defined as:

$$\begin{aligned} u_{acc}(\alpha, \beta) &= Acc_{(\alpha,\beta)}(psent), \\ u_{cov}(\alpha, \beta) &= Cov_{(\alpha,\beta)}(psent), \end{aligned} \tag{7}$$

where $Acc_{(\alpha,\beta)}(psent)$ and $Cov_{(\alpha,\beta)}(psent)$ are calculated in Equation (4) and Equation (5).

2) Game Implementation: In the game implementation phase, acc and cov take their strategies, and we calculate the payoffs under each α and β pair.

3) Result Acquisition: A strategy profile s of player o_i is Pareto efficient if for *no* strategy profile s' that [2]:

$$\begin{aligned} &\forall i \in \{1, ..., n\},\ o_i(s') \geqslant o_i(s), \\ \text{and } &\exists i \in \{1, ..., n\},\ o_i(s') > o_i(s) \end{aligned} \tag{8}$$

The definition of Pareto efficiency in Equation (8) refers to a situation where there does not exist an alternative feasible resource allocation (i.e., strategy profile) which can make some individuals better off (higher payoff) without making someone else worse off (lower payoff) [21]. Therefore, by applying Pareto efficiency, we can find an optimal result that both players' payoffs are highest possible.

Algorithm 1: Search for the optimal threshold pair

```
Input  : Payoff table $T$
Output: A row containing the optimal threshold pair
set stop criterion $P$ = 0.95;
    $row_i$: the i-th row in $T$;
    condition $Cn$ of finding the final $row_i$:
    $row_i$.accuracy $> P$ and $row_i$.coverage $> P$
for $P$ from 0.95 to 0 do
    for the first row to the last row in $T$ do
        if $row_i$ meets $Cn$ then
            output $row_i$;
        else
            decrease $P$ by 0.0001;
        end
    end
end
return $row_i$.threshold_pair: value of the attribute
  threshold_pair in $row_i$
```

In order to find the optimal result, we set a common goal for both players. If they cannot meet it, then we lower it until both players reach the goal. The game results are recorded in a payoff table which consists of three attributes: threshold pair, accuracy and coverage. The procedure of searching for the optimal result in the payoff table is explained in Algorithm 1. The condition of stop searching is to find one row whose accuracy and coverage are both above a stop criterion (i.e., the common goal). We set an initial stop criterion as 95%. If a row meets the stop criterion, then this row is our final result. If not, we lower the stop criterion value by 0.0001 and iterate until we get the final result, or the criterion reaches 0%.

IV. Experiment and Results

This section discusses the experimental datasets and the final results of our experiment.

A. Experimental Datasets

We have selected three experimental datasets in different sizes: Yelp polarity reviews dataset, IMDB movie reviews dataset, and Amazon product review dataset. All three datasets are standard datasets that have been widely experimented with in sentiment analysis research.

Yelp polarity reviews dataset is constructed and used as a text classification benchmark in [31]. It is initially extracted from the Yelp Dataset Challenge 2015 data and is currently available on the *https://course.fast.ai/* website. This dataset consists of 598,000 reviews, where the reviews with stars 1 and 2 are labeled as negative and 3 and 4 as positive.

IMDB movie reviews dataset is firstly introduced in [17], and publicly available on the website *https://ai.stanford.edu/*.

The authors have collected 50,000 movie reviews from the IMDB website and labeled negative reviews with scores no more than 4 out of 10; positive for those no less than 7 scores.

Amazon product reviews dataset is constructed and introduced in [5], and accessible on the *https://www.cs.jhu.edu/* website. The authors have selected four different product types: books, DVDs, electronics, and kitchen appliances, each containing 2,000 reviews. The reviews with more than 3 out of 5 ratings are labeled as positive; those less than 3 are negative. Our experiment combines the four product datasets into one larger dataset containing 8,000 reviews in total.

All three datasets are balanced datasets with exact half instances positive sentiment and half negative. Furthermore, the IMDB and Amazon datasets are highly polarized. In other words, the authors have excluded the neutral sentiment data based on the ratings or scores. However, the ratings or scores can only reveal a general orientation of a person's sentiment. The textual reviews following the ratings or scores can contain more detailed information regarding their opinions. Therefore, the proposed model may prove the existence of uncertainty and ambiguity in ratings-or-scores-based polarized datasets.

Moreover, the reviews in all three datasets are long texts containing multiple sentences. This allows us to avoid the situation where only a few words influence a review's sentiment polarity. Therefore, extracting more nouns, verbs, adjectives, and adverbs can capture more relevant information (i.e., sentiment polarity).

TABLE IV
A SUMMARY OF THE DATASETS IN THE EXPERIMENT

Dataset	Total Reviews	Positive Reviews	Negative Reviews	Average Word per Review (Cleaned)
Yelp	598,000	299,000	299,000	127.63
IMDB	50,000	25,000	25,000	222.03
Amazon	8,000	4,000	4,000	335.97

A summary of the datasets is shown in TABLE IV, including total reviews, positive and negative labeled reviews, as well as the average word count per review after the texts are cleaned (the detailed steps are discussed in the text cleaning subsection of Section III).

B. Game Outcomes

According to the defined strategies of the two players, we have obtained 121 threshold pairs from each dataset.

The possible threshold pairs and the corresponding payoffs (accuracy and coverage), are calculated and recorded in payoff tables. TABLE V shows part of the payoff table of Yelp dataset as an illustrative example. The payoff tables contain three columns: (α,β) threshold, accuracy and coverage.

C. Optimal Result Acquisition

After multiple trials of criteria based on Algorithm 1, we achieve the optimal threshold pairs of each dataset as shown in TABLE VI. In the Yelp dataset, when we set the criterion down to 80.77%, then the threshold pair (0.65, 0.40) pops up with the highest value pair of accuracy (80.77%) and coverage (81.51%). Similarly, in the IMDB dataset, the criterion as 77.85% gives us the optimal result: (0.65, 0.45) as thresholds with accuracy 78.09% and coverage 77.86%. In the Amazon dataset, when the criterion is 79.10%, the threshold pair (0.6, 0.35) achieves the balance between accuracy (79.10%) and coverage (80.04%).

TABLE V
PAYOFF TABLE OF YELP DATASET

	Threshold (Alpha, Beta)	Accuracy	Coverage
1	(1.0, 0.0)	1.0000	0.0031
2	(1.0, 0.05)	0.9762	0.0967
3	(1.0, 0.1)	0.9534	0.1651
4	(1.0, 0.15)	0.9371	0.2108
5	(1.0, 0.2)	0.9287	0.2293
...	...	...	...
117	(0.5, 0.3)	0.7939	0.8455
118	(0.5, 0.35)	0.7907	0.8707
119	(0.5, 0.4)	0.7844	0.9061
120	(0.5, 0.45)	0.7743	0.9512
121	(0.5, 0.5)	0.7617	1.0000

TABLE VI
EXPERIMENT RESULT FROM GTRS GAMES IN THREE DATASETS

Dataset	Optimal Threshold	Accuracy	Coverage
Yelp Dataset	(0.65, 0.40)	80.77%	81.51%
IMDB Dataset	(0.65,0.45)	78.09%	77.86%
Amazon Dataset	(0.6,0.35)	79.10%	80.04%

V. ANALYSIS AND COMPARISON

This section interprets the GTRS game results and compares the prediction performance with several other classification methods.

A. Result Interpretation

Based on the payoff tables acquired, we plot a line graph to illustrate the GRTS game results, Fig. 2 shows the result of Yelp dataset. The green (top) line and blue (bottom) line represent values of accuracy and coverage, respectively. The x-axis consists of all threshold pairs, and the y-axis represents accuracy and coverage values ranging from 0 to 1. The red (middle and vertical) lines highlight the optimal result, where the cross point's coordinate corresponds to the final threshold pair (x-axis) and the stop criterion for finding the final result (y-axis), that is, a common goal.

The figure illustrates how the payoffs of two players in the GTRS game vary under different strategies. As the strategy (pair of thresholds) changes, the overall trend of accuracy decreases while coverage increases. When two payoffs converge, they achieve a balanced state, and also their highest possible outcomes.

We may interpret the result of Yelp dataset in TABLE VI as follows: the highest accuracy achieved without compromising the data coverage is 80.77%, with 81.51% data to be classified as either positive or negative. On the other hand, obtaining a data coverage higher than 81.51% will compromise data accuracy.

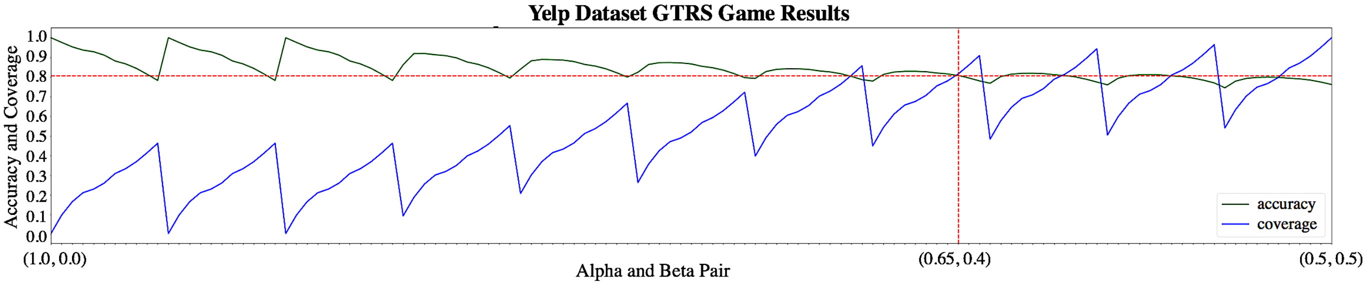


Fig. 2. Trends of accuracy and coverage rates in GTRS games

B. Comparison With Other Rough Set Models

To evaluate the proposed model's prediction performance, we adopt a 10-fold cross-validation method to train and test the model. In each fold, the dataset is divided 90% as training subset, and 10% as testing subset. The optimal thresholds are trained and obtained from each training subset. We then apply the same thresholds to divide *POS*, *BNG*, and *NEG* regions on the corresponding testing subsets, labeled as positive, neutral and negative sentiment, respectively. We average the accuracy of ten testing subsets as the model's prediction accuracy.

In GTRS games, two extreme cases in the game are worth discussion: Pawlak's rough sets and 0.5-probabilistic rough sets. The former model has (1,0) as the threshold pair, in which the classification is based on complete certainty leading to a 100% accuracy. The 0.5-probabilistic rough sets, on the contrary, has a threshold pair (0.5,0.5). In other words, only the equivalence classes with a conditional probability of 0.5 can be considered neutral. Therefore, this latter model has large data coverage. The two models are evaluated in the same method as the proposed model, and the comparison is shown in the first two rows of TABLE VII.

According to the evaluation results, Pawlak's rough set model usually has higher accuracy than the proposed model but compromises data coverage (as low as 0.36%, 0.52%, and 8.81%). The 0.5-probabilistic rough set model covers close to 100% of the data (100.00%, 98.83%, and 96.86%) but has less accuracy than the proposed model by 4.64%, 4.74%, and 4.19%. Therefore, we can conclude that the proposed model outperforms the other two rough set models and balances accuracy and coverage.

C. Comparison With Four Binary Classification Models

SVM, naïve Bayes, decision tree, and KNN are classification models that are widely applied in sentiment classification tasks [24]. In this experiment, we compare the evaluation results of these four models with the proposed model. Since all four models are binary classification, the coverage rates are 100%, which means all data are classified into either positive or negative. All three datasets apply the same preprocessing method and 10-fold cross-validation method as the proposed model. TABLE VII shows the comparison of prediction evaluation in detail. Among the four classification models, the highest accuracy rates in the three datasets are: in the Yelp dataset, decision tree achieves 76.11%; in the IMDB dataset, KNN achieves 72.62%; and in the Amazon dataset, SVM achieves 72.06%. Given that the proposed model has 80.76%, 77.06%, and 74.93% in three datasets, we conclude that our model outperforms the four classification models.

In summary, from the experimental result of the proposed model in three datasets, we have shown that the payoffs of two players are converging as the strategies change and finally achieve an optimal outcome. Our outcome has established a balance between accuracy and coverage compared with other rough set models, such as Pawlak's rough sets and 0.5-probabilistic rough sets. In comparing accuracy rates with other classic binary classification models, our proposed model also achieves better results.

Indeed, there are a few more resources that can be used to compare with our model. We will dedicate further work to evaluate other methods to extract words that contain a high contribution to sentiment scores for the text instances, such as SenticNet 6 [7].

VI. Conclusion

In conclusion, the complexity of texts and the uncertainty of opinions make sentiment analysis a challenging task. This paper has proposed a hybrid model using part-of-speech-based feature extraction and game-theoretic rough sets or GTRS to conduct sentiment classification on three different datasets. We extract nouns, verbs, adjectives, and adverbs from texts, then transform four part-of-speech groups into sentiment polarity scores using Valence Aware Dictionary for Sentiment Reasoning or VADER. We have significantly reduced the dimensionality of texts while preserving the sentiment information in the texts. Moreover, with GTRS, we have achieved highest possible results of accuracy and coverage to deal with the uncertainty of opinions, even on highly pre-polarized datasets (i.e., IMDB and Amazon). Comparing with three datasets' experiment results, the proposed model achieves higher coverage than Pawlak's rough sets by 80.97%, 78.41%, and 71.88%; and higher accuracy than 0.5-probabilistic rough sets by 4.64%, 4.74%, and 4.19%. Regarding the accuracy rates of four binary classification models, including SVM, naïve Bayes, decision tree, and KNN, the proposed model also outperforms by 4.65%, 4.44%, and 2.87%. The evaluation results show that the proposed model is promising to deal with the complexity and uncertainty in sentiment analysis.

TABLE VII
EVALUATION RESULTS COMPARISON

	Yelp Dataset		IMDB Dataset		Amazon Dataset	
Model	Accuracy	Coverage	Accuracy	Coverage	Accuracy	Coverage
Pawlak's	98.64%	0.36%	76.95%	0.52%	79.94%	8.81%
0.5-Probabilistic	76.12%	100.00%	72.32%	98.83%	70.74%	96.86%
SVM	73.68%	100.00%	70.42%	100.00%	72.06%	100.00%
Naïve Bayes	74.73%	100.00%	71.80%	100.00%	70.91%	100.00%
Decision Tree	76.11%	100.00%	72.58%	100.00%	71.74%	100.00%
KNN	72.54%	100.00%	72.62%	100.00%	71.83%	100.00%
Proposed Model	**80.76%**	**81.33%**	**77.06%**	**78.93%**	**74.93%**	**80.69%**

ACKNOWLEDGMENT

This research was partially supported by NSERC, the Natural Sciences and Engineering Research Council of Canada, and the Faculty of Graduate Studies and Research of the University of Regina.

REFERENCES

[1] M. S. Akhtar, A. Ekbal and E. Cambria, "How Intense Are You? Predicting Intensities of Emotions and Sentiments using Stacked Ensemble [Application Notes]," in *IEEE Computational Intelligence Magazine*, vol. 15, no. 1, pp. 64-75, Feb. 2020.

[2] K. R. Apt, "A Primer on Strategic Games," in Lectures in Game Theory for Computer Scientists, K. R. Apt and E. Grädel, Eds. Cambridge University Press, 2011, pp. 1–37.

[3] L. Augustyniak, P. Szymański, T. Kajdanowicz, and W. Tuliglowicz, "Comprehensive Study on Lexicon-based Ensemble Classification Sentiment Analysis," *Entropy*, vol. 18, p. 4, 2016.

[4] N. Azam and J. T. Yao, "Game-theoretic rough sets for recommender systems," *Knowledge-Based Systems*, vol. 72, pp. 96–107, 2014.

[5] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, Bollywood, Boomboxes and Blenders: Domain Adaptation for Sentiment Classification," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Jun. 2007, pp. 440–447.

[6] E. Cambria, "Affective Computing and Sentiment Analysis," in *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102-107, Mar.-Apr. 2016.

[7] E. Cambria, Y. Li, F. Z. Xing, S. Poria, and K. Kwok, "SenticNet 6: Ensemble Application of Symbolic and Subsymbolic AI for Sentiment Analysis," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 105–114.

[8] N. Chumerin and M. M. Van Hulle, "Comparison of Two Feature Extraction Methods Based on Maximization of Mutual Information," in *2006 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*, pp. 343-348, 2006.

[9] J. H. Dahooie, R. Raafat, A. R. Qorbani, and T. Daim, "An Intuitionistic Fuzzy Data-Driven Product Ranking Model using Sentiment Analysis and Multi-Criteria Decision-Making," *Technological Forecasting and Social Change*, vol. 173, p. 121158, 2021.

[10] A. Ghabayen and B. Ahmed, "Polarity Analysis of Customer Reviews Based on Part-of-Speech Subcategory," *Journal of Intelligent Systems*, vol. 29, Aug. 2019.

[11] E. Haddi, X. Liu and Y. Shi, "The Role of Text Pre-processing in Sentiment Analysis," *Procedia Computer Science*, vol. 17, pp. 26-32, 2013.

[12] J. Herbert and J. T. Yao, "Game-Theoretic Rough Sets," *Fundam. Inform.*, vol. 108, pp. 267–286, Jan. 2011.

[13] L. P. Hung and R. Alfred, "A Performance Comparison of Feature Extraction Methods for Sentiment Analysis," in *Asian Conference on Intelligent Information and Database Systems*, 2017, pp. 379–390.

[14] C. Hutto and E. Gilbert, "Vader: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2014, vol. 8, no. 1.

[15] P. R. Kroeger, Analyzing grammar: An introduction. Cambridge University Press, 2005.

[16] B. Liu, Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge University Press, 2015.

[17] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng and C. Potts, "Learning Word Vectors for Sentiment Analysis", in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Jun. 2011, pp. 142–150.

[18] C. Nicholls and F. Song, "Improving sentiment analysis with Part-of-Speech weighting," in *2009 International Conference on Machine Learning and Cybernetics*, 2009, vol. 3, pp. 1592–1597.

[19] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, Jul. 2002, pp. 79–86.

[20] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data. Springer Netherlands, 1991.

[21] J. E. Stiglitz, "Pareto Optimality and Competition," The Journal of Finance, vol. 36, no. 2, pp. 235–251, 1981.

[22] L. Sun, J. Guo, and Y. Zhu, "A Multi-Aspect User-Interest Model Based on Sentiment Analysis and Uncertainty Theory for Recommender Systems," *Electronic Commerce Research*, vol. 20, no. 4, pp. 857–882, 2020.

[23] Z. Wang, S.-B. Ho, and E. Cambria, "Multi-Level Fine-Scaled Sentiment Sensing with Ambivalence Handling," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 28, no. 4, pp. 683–697, 2020.

[24] G. Wang, J. Sun, J. Ma, K. Xu, and J. Gu, "Sentiment classification: The contribution of ensemble learning," *Decision support systems*, vol. 57, pp. 77–93, 2014.

[25] Y. Y. Yao, "An Outline of a Theory of Three-Way Decisions," in *Rough Sets and Current Trends in Computing*, 2012, pp. 1–17.

[26] Y. Y. Yao, S. Greco, and R. Słowiński, "Probabilistic Rough Sets," in *Springer Handbook of Computational Intelligence*, J. Kacprzyk and W. Pedrycz, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, pp. 387–411.

[27] W. Zhang, T. Yoshida, and X. Tang, "A Comparative Study of TF* IDF, LSI and Multi-words for Text Classification," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2758–2765, 2011.

[28] Z. Zhang and R. Wang, "Applying Three-way Decisions to Sentiment Classification with Sentiment Uncertainty," in *Rough Sets and Knowledge Technology*, 2014, pp. 720–731.

[29] Y. Zhang, P. Liu and J. T. Yao, "Three-way Email Spam Filtering with Game-theoretic Rough Sets," *2019 International Conference on Computing, Networking and Communications (ICNC)*, 2019, pp. 552-556.

[30] Y. Zhang and J. T. Yao, "Multi-criteria Based Three-Way Classifications with Game-Theoretic Rough Sets," in *Foundations of Intelligent Systems*, 2017, pp. 550–559.

[31] X. Zhang, J. Zhao, and Y. LeCun, "Character-Level Convolutional Networks for Text Classification," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, 2015, pp. 649–657.

[32] Y. Zhou, Y. Zhang, and J. T. Yao, "Satirical News Detection with Semantic Feature Extraction and Game-Theoretic Rough Sets," in *Foundations of Intelligent Systems*, 2020, pp. 123–135.