January 2019

Methods For Classification Of Consumer Review Into The Ones Written Before Or After The Product Purchase

Md Mehedi Hasan
Wayne State University, mehedi2003@gmail.com

METHODS FOR CLASSIFICATION OF CONSUMER REVIEW INTO THE ONES WRITTEN BEFORE OR AFTER THE PRODUCT PURCHASE

by

MD MEHEDI HASAN

THESIS

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

2019

MAJOR: COMPUTER SCIENCE

Approved By:

Advisor Date

DEDICATION

This thesis is dedicated to Allah and my parents, for all their love, patience, kindness and support.

I also dedicate this to my daughter and wife who are my everything and always been my greatest inspiration.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisor, Dr. Alexander Kotov, who generously offered his always wise guidance. I am very thankful for his support, his patience, and his time. Without his guidance and persistent help, this thesis is not possible. I also want to thank my committee members for their time and support.

In addition, I would like to thank my colleagues and friends at Textual Data Analytics (TEANA) Lab: Fedor, Saeid and Diana for their help and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

The content posted on online consumer review platforms contains a wealth of information, which besides positive and negative judgments about product features and services, often includes specific suggestions for their improvement and root causes for customer dissatisfaction. Such information, if accurately identified, could be of immense value to businesses. Although previous research on consumer review analysis has resulted in accurate and efficient methods for classifying reviews according to the overall sentiment polarity [26], segmenting reviews into aspects and estimating the sentiment score of each aspect [33], as well as summarizing both aspects and sentiments towards them [16, 30, 31, 29], more focused types of review analysis, such as detecting the intent or the timing of reviews, are needed to better assist companies in making business decisions. One such problem is separating the reviews (or review fragments) written by the users after purchasing and using a product or a service (which we henceforth refer to as "post-purchase" reviews) from the reviews that are written by the users, who shared their expectations or results of research before purchasing and using a product (which we henceforth refer to as "pre-purchase" reviews).

We hypothesize that effective separation of these two types of reviews (or review fragments) can allow businesses to better understand the aspects of products and services, which the customers are focused on before and after the purchase and tailor their marketing strategies accordingly. It can also allow businesses to measure the extent to which customer expectations are met by their existing products and services. Furthermore, "post-purchase" reviews, particularly the negative ones, can be considered as "high priority" reviews since they provide customer feedback, which needs to be immediately acted upon by manufacturers. Such feedback typically contains reports of malfunctions, as well as poor performance of products that are already on the market. Pre-purchase reviews, on the other hand, are likely to be written for ex-

pensive products that are major purchasing decisions and require extensive research prior to purchase (e.g. cameras, motorcycles, boats, cars, etc.). Such products often have a community of enthusiasts, who often post reviews of the product models they have only heard or read about.

In this work, we introduce a novel text classification problem of separating pre-purchase from post-purchase consumer review fragments. While, in some cases, the presence of past tense verb(s) or certain keywords in a given review fragment provides a clear clue about its timing with respect to purchasing (e.g. "excellent vehicle, great price and the dealership provides very good service"), other cases require distinguishing subtle nuances of language use and making inferences. For example, although the past tense verbs in "The new Ford Explorer is a great looking car. I heard it has great fuel economy for an SUV" and "so far this is the best car I tested" indicate prior experience, these review fragments are written by the users, who didn't actually purchase the products. Despite an overall positive sentiment of these review fragments, they provide no specific information to the manufacturer about how these cars can be improved. On the other hand, while the fragment "If I could, I would have two" refers to the future, it is clearly post-purchase.

To address the proposed problem, we evaluate the effectiveness of the features based on dictionaries and part-of-speech (POS) tags, in addition to the lexical ones. The key contributions of this work are two-fold:

1. We introduce a novel review analysis problem and provide a publicly available gold standard to evaluate the approaches to solve it;

2. We experimentally demonstrate that using both dictionary and POS pattern-based features allows to improve the performance of classifiers for this problem relative to using either of these feature types or lexical features alone.

The rest of the thesis is structured as follows. In chapter 2, we provide background

about machine learning and natural language processing. Chapter 3 provides an overview of previous related work. Chapter 4 describes the details of the experimental setup, while chapter 5 presents our main results. Chapter 6 discusses the results and chapter 7 summarizes the key contributions of this work and outlines future directions.

CHAPTER 2 MACHINE LEARNING AND NATURAL LANGUAGE PROCESSING FUNDAMENTALS

2.1 Machine Learning (ML)

Machine learning (ML) is an algorithm that provides the capability of a computer system to learn on the basis of their own previous experience and act accordingly without being explicitly programmed. Over the last decade, machine learning has been applied in a wide range of applications. One of the most well-known examples is facebook and google news feed. Beside that machine learning is also utilized in virtual personal assistants, GPS navigation, effective web search, self-driving cars, speech recognition, cancer prognosis and prediction [21, 5, 15, 19]. It is so widely used in the daily life that we probably used a dozen of machine learning based application per day without knowing it. Machine learning is also pervasive in enterprise applications such as online product recommendations or fraud detection [20, 23]. Sometimes, a customer service chatbot is integrated into e-commerce websites to assist their customer for shopping [8]. Customer relationship management (CRM) systems use machine learning models to analyze email and inform sales team members to respond to the most important messages first. More advanced systems can even recommend potentially effective responses. ML models also used to analyze customers feedback and suggestion and rely on this knowledge to improve their products [9]. Machine learning algorithms are categorized into supervised, unsupervised, semi-supervised and reinforcement learning.

2.1.1 Supervised Learning

Supervised learning algorithm allows a computer application to learn from the given input and desired output and the algorithm will apply this knowledge to make a prediction for new data. Therefore, human intervention is required to collect the labeled data. Data scientist or data analyst determine which variables, or features, the model should analyze and use to develop predictions. All classification and regres-

sion algorithms such as linear regression, logistic regression, support vector machine, decision tree and naive bayes are examples of supervised learning.

2.1.2 Unsupervised Learning

Unsupervised algorithms do not need labeled data for training instead the algorithm is used to discover the underlying structure of the data. Since the outcomes are unknown, evaluation of unsupervised machine learning algorithms is more challenging. A typical example of unsupervised machine learning algorithms is k-means clustering, hierarchical clustering and dimensionality reduction, which were utilized by many applications include anomaly detection and feature learning [35, 6].

2.1.3 Semi-supervised Learning

Labeled data are expensive because it requires an expert to manually labeling the data, which is a tedious and time-consuming process. In addition, too much labeling can impose human biases on the model. On the other hand, unlabeled data are cheaper and most of the real world data are unlabeled. To utilize a huge amount of unlabeled data, a combination of supervised and unsupervised learning is utilized known as semi-supervised learning. For semi-supervised learning, we need a small amount of labeled data for a large amount of unlabeled data. For that reason, semi-supervised learning is often utilized for webpage classification, speech recognition, or even for genetic sequencing [32, 24].

2.1.4 Reinforcement Learning

Reinforcement learning is a learning algorithm, in which an agent learns in a new environment by taking action and seeing the results. When the agent receives a reward we measure the agents action as success otherwise penalize it. Reinforcement learning applied on many applications include robotics, traffic light control and optimizing chemical reactions [34, 18, 2].

2.2 Training Algorithms

A learning algorithm is used to train an ML model to learn from training data. The training dataset contains the label of the input data, which is known as a target or target attribute. The main objective of the training algorithm is to learn a function from given training data by minimizing loss, where mean squared error (MSE) and cross entropy often used as loss functions for regression and classification task, respectively. Model parameters are estimated by minimizing the loss during the training. Example of some training algorithms is gradient descent, mini-batch gradient descent and online learning [13].

2.2.1 Underfitting

A simple model often suffers from high bias by the assumption of simplicity is called underfitting. In this case, the trained model doesn't learn enough correlations between independent variables or predictors and dependent variable or target. It does not fit the training data as well. To resolve this issue, a complex function should be learned such as higher degree polynomial should be learned instead of a simple linear model.

2.2.2 Overfitting

When a trained model finds a correlation from training dataset that may not exist, this would be called overfitting. In this case, the model achieves a high variance to learn a complex function by adopting the given training inputs. The possible solution to this problem is to use more data or use regularization so that the model will not be able to learn a complex function from the training dataset.

2.3 Regularization

Regularization is an important concept in machine learning used to solve the overfitting problem. Regularizations techniques reduce the generalization error by fitting a function appropriately on the given training dataset. Without a substantial

increase in its bias, this technique significantly reduces the variance of the model. Regularization is applied by adding an additional penalty term in the error function. The additional term controls the model function such that the coefficients dont take very large values.

2.3.1 L2-Regularization or Ridge Regression

The Ridge regression is also known as L2-regularization uses $L2$ norm for regularization. Ridge regression adds a squared magnitude of coefficient as a penalty term to the loss function. L2-regularization corresponding to Gaussian prior and inclines to spread error among all variables.

2.3.2 L1-Regularization or Lasso Regression

The lasso regression is also known as L1-regularization uses $L1$ norm for regularization. L1-regularization corresponds to Laplacian prior and spreads error among a few independent variables. Lasso regression adds the absolute value of the magnitude of coefficient as a penalty term to the loss function. The main difference between ridge and lasso regression is a shape of the constraint region. The main advantage of using lasso regression is that it shrinks the less important features coefficient to zero. Therefore, it works well for feature selection and widely used when we have a huge number of features.

2.4 Evaluation Metrics

We report standard metrics of accuracy, precision, recall and F1-measure to evaluate the performance of the classifiers.[1] The results are reported based on k-fold cross-validation (one fold was used as a test set and the remaining k-1 folds were used as a training set) and weighted macro-averaging over the folds. In the following section, accuracy, precision, recall and F1-measure are defined in terms of true positive (TP), false positive (FP) and false negative (FN). A true positive (TP) was counted when the method correctly classified an instance into its actual class; a false positive

(FP) was counted for a class when the method incorrectly classified an instance into that class; a false negative (FN) for an actual class of instance was counted when the method incorrectly classified the instance into other class.

2.4.1 Accuracy

Accuracy is the number of instances correctly predicted divided by the total number of predictions made. Accuracy is not enough to measure the performance of a model because it is misleading for an imbalance dataset such as cancer dataset where only a small percentage of patients might have cancer. Therefore, additional measures are required to evaluate the performance of a classifier.

2.4.2 Precision

The precision of a class was defined as the ratio of the numbers of correctly classified instances and the total number of instances identified as belonging to that particular class by the classifier. Precision is the measure of relevant or exactness and this metric is matters for web search results and spam filtering. The goal of the spam filtering algorithm is to minimize the number of reals emails that are classified as spam.

$$Precision = \frac{TP}{TP + FP}$$

2.4.3 Recall

Recall of a class was defined as the ratio between the numbers of correctly classified instances and the total number of instances of that particular class in the gold standard. The recall is a matter when we don't care about false positives but really want to hit every single positive case. For example, if a patient has some of the cancer symptoms and the prediction model says that the patient has the possibility of having cancer and need a followup test. After the blood test, if we don't find a positive result of cancer then we only lose some money. But if we don't take it seriously, the patient

could have cancer and be dead in a month.

$$Recall = \frac{TP}{TP + FN}$$

2.4.4 F1-Measure

F1-measure is used when precision and recall both are equally important. F1-measure is computed as the harmonic mean of precision and recall. A good F1-measure indicates that you have both low false positives and low false negatives in your predictions.

$$F1 - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

2.4.5 K-folds Cross-validation

Machine learning models were usually evaluated by dividing the original dataset into a training dataset and test dataset. The training dataset is used for training the model whereas test dataset is used for evaluation of the models performance. The main problem with this evaluation technique is that it does not give an indication of how well the learner will generalize to an unseen data set. Cross-validation is a technique to evaluate predictive models by partitioning the original sample into k equal size subsamples called fold. The first fold is kept for testing and the model is trained on k-1 remaining folds. The process is repeated k times and each time different fold or a different group of data points are used for validation. For classification problems, one typically uses stratified k-fold cross-validation, in which the folds are selected so that each fold contains approximately the same proportions of class labels.

2.5 Natural Language Processing (NLP)

Computers work great with standardized and structured data like database tables and financial records. But humans communicate using words, a form of unstructured data. The customer provides their feedback about a product in social media or

product's websites as a form of natural language. Popular products have thousands of reviews which are hard for a human to read all these reviews, required thousands of hours. Therefore, an efficient NLP method required to perform large-scale analysis using natural language processing. NLP made this task easier for the company to identify a loyal customer and product issues quickly so that they can create data-driven strategies.

2.5.1 Part-of-speech (POS) Tagging

Part-of-speech (POS) tagging is a process of labeling words with their appropriate part-of-speech, where a word is labeled as one of the eight main parts of speech: noun, pronoun, verb, adverb, adjective, preposition, conjunction and interjection. POS tags are useful in various NLP tasks including text to speech conversion. If we know the verb of a sentence then we can estimate that what action(s) the sentence is talking about, and many NLP systems concentrate on the POS tags when trying to understand what a text is about.

CHAPTER 3 RELATED WORK

Although consumer reviews have been a subject of many studies over the past decade, a common trend of recent research is to move from detecting sentiments and opinions in online reviews towards the broader task of extracting actionable insights from customer feedback. One relevant recent line of work focused just on detecting wishes [14, 28] in reviews or surveys. In particular, Goldberg et al. [14] studied how wishes are expressed in general and proposed a template-based method for detecting the wishes in product reviews and political discussion posts, while Ramanand et al. [28] proposed a method to identify suggestions in product reviews. Moghaddam [22] proposed a method based on distant supervision to detect the reports of defects and suggestions for product improvements in online reviews.

Other non-trivial textual classification problems have also been recently studied the literature. For example, Bergsma et al. [4] used a combination of lexical and syntactic features to detect whether the author of a scientific article is a native English speaker, male or female, or whether an article was published in a conference or a journal, while de Vel et al. [10] used style markers, structural characteristics and gender-preferential language as features for the task of gender and language background detection.

CHAPTER 4 METHODS

4.1 Data Collection

To create the gold standard for experiments in this work[1], we collected the reviews of all major car makes and models released to the market in the past 3 years from MSN Autos[2]. Table 4.1 and Figure 4.1 provide the examples of car reviews given by their customer. We segmented the reviews into individual sentences, removed punctuation except exclamation (!) and question (?) marks (since [3] suggest that retaining them can improve the results of some classification tasks), and annotated the review sentences using Amazon Mechanical Turk. In order to reduce the effect of annotator bias, we created 5 HITs (Human Intelligence Tasks) per each label and used the majority voting scheme to determine the final label for each review sentence. In total, the gold standard consists of 3983 review sentences. Table 4.2 shows the distribution of these sentences over classes. We used unigram bag-of-words lexical feature representation for each review fragment as a baseline, to which we added four binary features based on the dictionaries and four binary features based on the POS tag patterns manually compiled as described in Section 4.2.

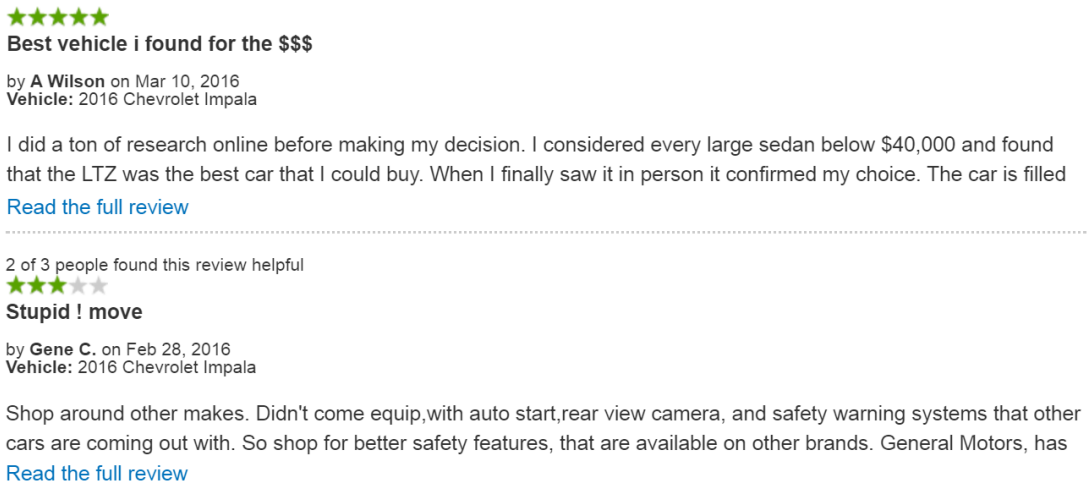
4.2 Features: Lexical Features, Dictionaries and POS Patterns

Lexical features were derived from a unigram bag-of-words representation of consumer reviews. On the other hand, each of the dictionaries contains the terms, which represent a particular concept related to the product, such as negative emotion, ownership, satisfaction, etc. To create the dictionaries, we first came up with a small

[1] gold standard and dictionaries are available at `http://github.com/teanalab/prepost`
[2] `http://www.msn.com/en-us/autos`

Table 4.1: Examples of customer reviews before the product purchase and after the product purchase

Customer Review	Class
"would not buy this, I would stick to the Ford F-150"	pre-purchase
"the best truck i have ever owned"	post-purchase



★★★★★
Best vehicle i found for the $$$

by **A Wilson** on Mar 10, 2016
Vehicle: 2016 Chevrolet Impala

I did a ton of research online before making my decision. I considered every large sedan below $40,000 and found that the LTZ was the best car that I could buy. When I finally saw it in person it confirmed my choice. The car is filled
Read the full review

2 of 3 people found this review helpful
★★★☆☆
Stupid ! move

by **Gene C.** on Feb 28, 2016
Vehicle: 2016 Chevrolet Impala

Shop around other makes. Didn't come equip,with auto start,rear view camera, and safety warning systems that other cars are coming out with. So shop for better safety features, that are available on other brands. General Motors, has
Read the full review

Figure 4.1: Example of customer reviews about the vehicle

Table 4.2: Distribution of classes in experimental dataset

Class	# samples	Percentage
Pre-purchase	2122	53.28 %
Post-purchase	1861	46.72 %
Total	3983	100 %

set of seed terms, such as “buy”, “own”, “happy”, “warranty”, that capture the key lexical clues related to the timing of review creation regardless of any particular type of product. Then, we used on-line thesaurus[3] to expand the seed words with their synonyms and considered each resulting set of words as a dictionary.

Using a similar procedure, we also created a small set of POS tag-based patterns that capture the key syntactic clues related to the timing of review creation with respect to the purchase of a product. For example, the presence of combinations of

[3] `http://www.thesaurus.com`

Table 4.3: Dictionaries with associated words and phrases

Dictionary	Words
OWNERSHIP	own, ownership, owned, mine, individual, personal, etc.
PURCHASE	buy, bought, acquisition, purchase, purchased, etc.
SATISFACTION	happy, cheerful, contented, delighted, glad, etc.
USAGE	warranty, guarantee, guaranty, cheap, cheaper, etc.

Table 4.4: Part-of-speech (POS) pattern features with example of customer reviews

Pattern type	**Patterns**	**Example**
OWNERSHIP	**PRP$ CD**, PRP VBD, VBZ PRP$, VBD PRP$	this is **my third** azera from 2008 to 2010 until now a 2012
QUALITY	JJ, JJR, **JJS**	it is definitely the **best** choice for my family
MODALITY	**PRP MD**, IN PRP VBP	buy one **you will** love
EXPERIENCE	VBD, **VBN**	i have **driven** this in the winter and the all wheel drive model

possessive pronouns and cardinal numbers (pattern "PRP$ CD", e.g. matching the phrases "my first", "his second", etc.), personal pronouns and past tense (pattern "PRP VBD", e.g. matching "I owned") or modal (pattern "PRP MD", e.g. matching "I can", "you will", etc.) verbs, past participles (pattern "VBN", e.g. matching "owned or driven"), as well as adjectives, including comparative and superlative (patterns "JJ", "JJR" and "JJS") indicates that a review is likely to be post-purchase. More examples of dictionary words and POS patterns are provided in Tables 4.3 and 4.4.

4.3 Classifiers

We used Naive Bayes (NB), Support Vector Machine (SVM) with linear kernel implemented in Weka machine learning toolkit[4], as well as L2-regularized Logistic Regression (LR) implemented in LIBLINEAR[5][12] as classification methods.

4.3.1 Naive Bayes (NB)

Naive Bayes (NB) is a popular probabilistic method [17, 25] for text classification because of its robustness and relative simplicity. Naive Bayes assumes conditional independence of words in a textual fragment given its label. Although independence does not generally hold in practice. For a given input $x = x_1, x_2, ..., x_n$, the output class y is computed with features x_1 through x_n and classes c_1 through c_k (in our case $k = 2$) by the following formula:

[4] `http://www.cs.waikato.ac.nz/ml/weka`
[5] `http://www.csie.ntu.edu.tw/~cjlin/liblinear`

$$P(c_i|x_1, ..., x_n) \propto P(x_1, ..., x_n|c_i)P(c_i)$$

$$P(c_i|x_1, ..., x_n) \propto P(c_i) \prod_{j=1}^{n} P(x_j|c_i)$$

$$y = \arg\max_{c_i} P(c_i) \prod_{j=1}^{n} P(x_j|c_i)$$

Naive Bayes has demonstrated competitive performance over the years relative to other more sophisticated classifiers. Experimental results reported in this paper were obtained using standard implementations of multinomial Naive Bayes algorithms provided by the Weka toolkit[6]. Unlike binomial Naive Bayes, which only takes into account the presence or absence of a word from the collection vocabulary in a textual fragment for its classification, multinomial Naive Bayes classifier also takes into account the number of times words from the collection vocabulary occur in the fragment.

4.3.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) belongs to a family of generalized linear binary classifiers [7, 11], which tends to maximize the geometric margin of the classes and minimizes the empirical classification error. SVM uses kernel tricks, maps the low dimensional input feature vector into a higher dimensional space and finds a hyperplane that separates the samples into two classes in such a way that the margin between the closest samples in each class is maximized. Figure 4.2 illustrates that SVM finds the widest possible separating margin by allowing the misclassification of one sample. Equation of separating hyperplane is denoted by the following equation:

$$w^T x + b = 0$$

Here, b, x, and w represent the bias, training examples, and normal to the hy-

[6] `http://www.cs.waikato.ac.nz/ml/weka/`

perplane, respectively. Open-source implementation of SVM with a linear kernel in publicly available LibSVM[7] package was used for experiments reported in this work.

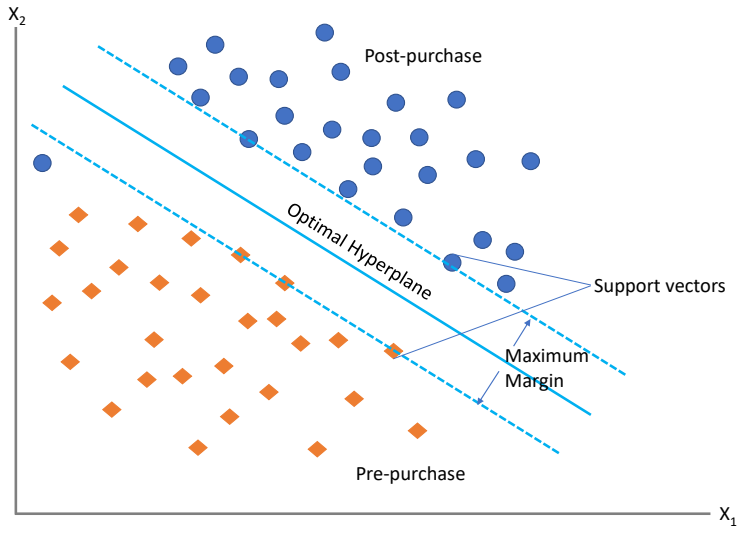


Figure 4.2: Optimal hyperplane of SVM to maximize the margin between two classes

4.3.3 Logistic Regression (LR)

Logistic Regression (LR) can in many ways be seen to be similar to linear regression [27] which models the relationship between one dependent or target variable and one or multiple independent variables. Linear regression also allows us to look at the fit of the model as well as at the relevance of the relationships that we are modeling. However, the underlying principle of binomial logistic regression and its statistical calculation are quite different from linear regression. Ordinary least square is used to find the best fitting line for a linear regression model. While linear regression estimates the optimal coefficients that predict the change in the value of the independent variable results in one unit change in the dependent variable, LR estimates the probability of an event. The sigmoid function is used to map predicted values into probabilities between 0 and 1. For given training samples X and optimal coefficients

[7] `https://www.csie.ntu.edu.tw/~cjlin/libsvm/`

β_0 and β, we used the following formula of the hypothesis $h\theta(X)$ for logistic regression model:

$$h\theta(X) = \frac{1}{1 + e^{-(\beta_0 + \beta X)}}$$

Figure 4.3 demonstrates that LR is more sensitive to outliers and tends to maximize posterior class probability. In our experiment, we used L2-regularized logistic regression implemented in LIBLINEAR[8] [12] as classification methods.

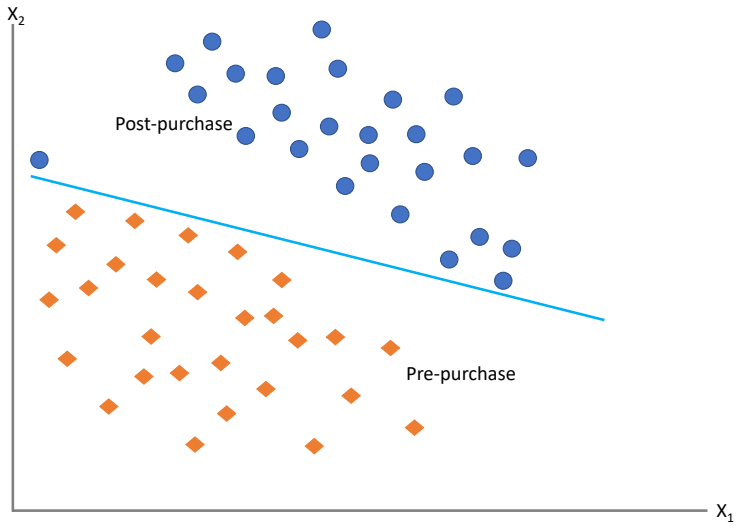


Figure 4.3: LR decision boundary that maximizes the posterior class probability

[8] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

CHAPTER 5 RESULTS

5.1 Classification of post-purchase vs. pre-purchase reviews using only lexical features

Performance of different classifiers for the task of separating post-purchase from pre-purchase reviews using only lexical features according to the standard performance metrics is shown in Table 5.1.

Table 5.1: Performance of different classifiers using only lexical features. Highest value for each metric across all models is highlighted in bold.

Method	Precision	Recall	F1-measure	Accuracy
SVM	**0.734**	0.724	0.717	0.724
LR	0.729	**0.726**	**0.722**	**0.726**
NB	0.703	0.704	0.702	0.704

Several observations can be made based on the results in Table 5.1. First, Regularized Logistic Regression outperforms SVM and Naive Bayes in terms of all performance metrics except precision. LR achieved 0.722 F1-Score with 0.729 precision and 0.726 recall. Second, Naive Bayes shows the lowest performance among all classifiers. Third, LR and SVM both provide 2.0-2.2% more accurate results than Naive Bayes for this classification task, and have similar accuracy (72.60% and 72.40% respectively) for this task. Although SVM demonstrates similar performance as LR, it achieves the highest precision for the task of classifying pre-purchase reviews from post-purchase reviews.

5.2 Classification of post-purchase vs. pre-purchase reviews using a combination of lexical, dictionary and POS pattern features

In the second set of experiments, to determine the relative influence of different features types, we obtained the performance of SVM, NB and LR methods in conjunction with the following features: i) combination of lexical and POS pattern features ii) combination of lexical features with dictionary features iii) combination of all three feature types (lexical, dictionary and POS pattern features). Summary of

the performance of classifiers with a combination of lexical and POS pattern features is provided in Table 5.2.

Table 5.2: Performance of different classifiers with combination of lexical and POS pattern features. The improvement in percentage is relative to using only lexical features by the same classifier. The highest value and largest improvement of each performance metric for a particular feature is highlighted in boldface and italic, respectively.

Method	**Prec.**	**Recall**	**F1-measure**	**Accuracy**
SVM	**0.733**(-0.17%)	0.727(+0.41%)	0.722(+0.70%)	0.727(+0.41%)
LR	**0.733**(+0.55%)	**0.730**(+0.55%)	**0.727**(+0.70%)	**0.730**(+0.55%)
NB	0.709(*+0.85%*)	0.710(*+0.85%*)	0.709(*+1.0%*)	0.710(*+0.85%*)

Comparing the influence of POS pattern features, several conclusions can be made. First, LR achieved the highest F1-score of 72.7% with 73.3% precision and 73% recall among all classification methods using POS pattern features with lexical features. Second, NB achieved better improvement relative to the lexical baseline than both SVM and LR, when POS pattern-based features were used. Third, the precision of SVM decreased by 0.17% while it's recall and F1-measure increased by 0.41% and 0.7%, respectively.

Table 5.3: Performance of different classifiers with combination of lexical and dictionary features. The improvement in percentage is relative to using only lexical features by the same classifier. The highest value and largest improvement of each performance metric for a particular feature is highlighted in boldface and italic, respectively.

Method	**Prec.**	**Recall**	**F1-measure**	**Accuracy**
SVM	**0.750**(*+2.18%*)	**0.741**(*+2.35%*)	**0.735**(*+2.51%*)	**0.741**(*+2.35%*)
LR	0.740(+1.51%)	0.736(+1.38%)	0.733(+1.52%)	0.736(+1.38%)
NB	0.713(+1.42%)	0.714(+1.42%)	0.713(+1.57%)	0.714(+1.42%)

Table 5.3 illustrates the influence of dictionary features on SVM, LR and NB methods. Results indicate that dictionary feature is more influential on SVM model, achieved the highest performance in terms of all performance metrics. SVM also demonstrates the highest improvement of model performance among all classifiers, improved 2.18%, 2.35%, 2.51% and 2.35% precision, recall, F1-measure and accuracy, respectively, when dictionary feature is used in addition to lexical feature. LR

performed better than NB classifier while it shows lower performance improvement compared to NB when dictionary features were used in conjunction with lexical features.

Table 5.4: Performance of different classifiers with combination of lexical, dictionary and POS pattern features. The improvement in percentage is relative to using only lexical features by the same classifier. The highest value and largest improvement of each performance metric for a particular feature is highlighted in boldface and italic, respectively.

Method	**Prec.**	**Recall**	**F1-measure**	**Accuracy**
SVM	**0.752**(*+2.45%*)	**0.743**(*+2.62%*)	**0.738**(*+2.93%*)	**0.743**(*+2.62%*)
LR	0.745(+2.19%)	0.741(+2.07%)	0.738(+2.22%)	0.741(+2.07%)
NB	0.717(+1.99%)	0.718(+1.99%)	0.717(+2.14%)	0.718(+1.99%)

Table 5.4 illustrates the influence of both dictionary and POS pattern features on different classification methods. We observe that SVM achieves the highest performance among all classifiers in terms of precision (0.752), recall (0.743) and accuracy (0.743) when a combination of lexical, POS and dictionary-based features was used. NB shows the lower performance than both SVM and LR, which is consistent with results in Table 5.2 and 5.3.

CHAPTER 6 DISCUSSION

Nowadays, most of the organizations are data-driven means all executive decisions should be made based on consumer data because it is important for the company to engage with customers emotionally by understanding consumer behavior and consumer insights. Online reviews and social media are a great place to collect consumer feedback. A text analytics solution is used to analyze online reviews of different brands and compare them against each other. This analysis will be invaluable for organizations to determine their marketing strategy or improving their product quality for their clients. Our study is important because this is the first step to discover actionable insights from consumer reviews. In this study, pre-purchase consumer reviews are separated from post-purchase consumer reviews so that further analysis can be done on consumer reviews of different groups.

Experimental results indicate that SVM is the best model among all machine learning methods considered for this study when all features are utilized together. However, using only lexical feature or POS pattern-based features in addition to lexical ones, LR achieves the highest performance in terms of all metrics and resulted in the highest improvement for NB classifier. On the other hand, a combination of lexical, dictionary and POS pattern-based features is more effective for SVM than for both NB and LR. Overall, experimental results presented above indicate that dictionary and POS pattern features allow to improve the performance of all classifiers for the task of separating pre-purchase from post-purchase review fragments relative to using only lexical features.

As a result, we noticed the trend of increasing performance for additional feature. Figures 6.1, 6.2, 6.3 and 6.4 illustrate the performance of SVM, LR and NB models in terms of precision, recall, accuracy and F1-measure, respectively, when different combination of features are utilized.

As follows from Figure 6.1, the influence of different features in precision on various

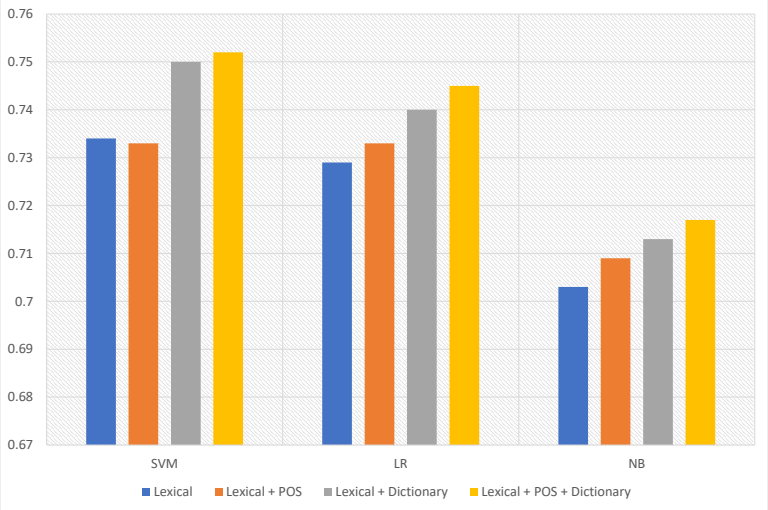


Figure 6.1: Performance of SVM, LR and NB models in terms of precision when different combination of features are utilized

classifiers are consistent except SVM when lexical features are used in combination with POS pattern feature. Although precision decreases for POS pattern features, SVM achieves the highest precision when dictionary and POS pattern features are used in addition to lexical features.

Figure 6.2 demonstrates the recall of SVM, LR and NB classifiers. Unlike precision, recall is improved for all classifiers with all combination of dictionary, POS pattern and lexical features. Similar to precision, SVM achieves the highest recall while NB attains it's lowest recall for all combination of input features. We also observed the same trends in accuracy and F1-measure which provide the evidence for the robustness of our model for the task of separating pre-purchase consumer reviews from post-purchase consumer reviews.

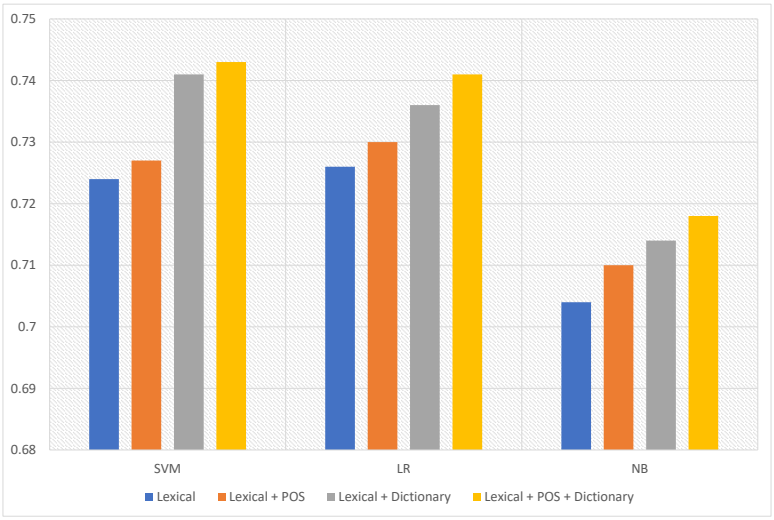


Figure 6.2: Performance of SVM, LR and NB models in terms of recall when different combination of features are utilized

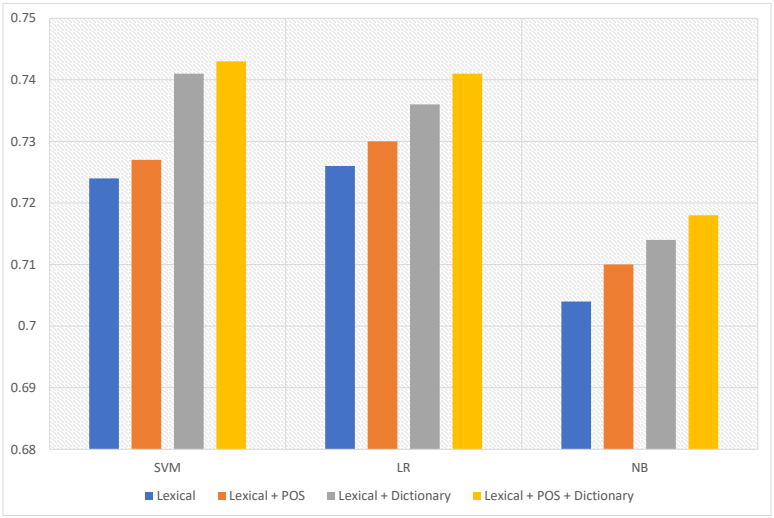


Figure 6.3: Performance of SVM, LR and NB models in terms of accuracy when different combination of features are utilized

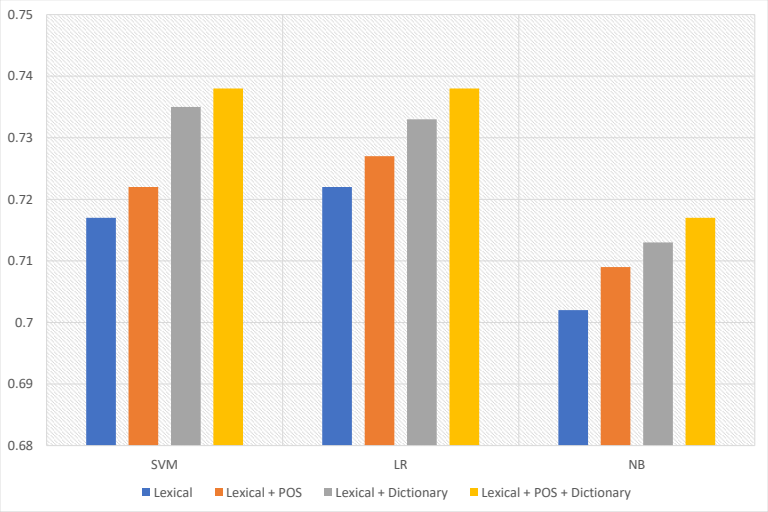


Figure 6.4: Performance of SVM, LR and NB models in terms of F1-measure when different combination of features are utilized

CHAPTER 7 CONCLUSION

In this study, we introduced a novel task of identifying post-purchase from pre-purchase reviews. This task is practically important for companies and constitutes an important step towards extracting actionable insights from online consumer reviews. We also experimentally demonstrated that combining lexical features with dictionary and POS pattern features allows improving the accuracy of all classification models that we examined. As future work, we propose to incorporate more information about the user, as features into classification tasks. Also, we would like to investigate this classification task with other state-of-the-art machine learning methods which can yield better results. Since the methods presented in this paper can be applied easily to any customer reviews, we plan to evaluate it using large scale datasets in order to directly compare the results with our present work.

APPENDIX

Gold Standard: a term used to describe a collection of a labeled dataset which has been manually labeled by the experts.

State-of-the-art: the most recent or latest version of a particular technology. State-of-the-art machine learning methods refer to the best available machine learning methods developed using modern techniques and technologies.

Prec.: Precision

REFERENCES

[1] Aas, K., and Eikvil, L. Text categorisation: A survey, 1999.

[2] Arel, I., Liu, C., Urbanik, T., and Kohls, A. Reinforcement learning-based multi-agent system for network traffic signal control. *IET Intelligent Transport Systems 4*, 2 (2010), 128–135.

[3] Barbosa, L., and Feng, J. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd international conference on computational linguistics: posters* (2010), Association for Computational Linguistics, pp. 36–44.

[4] Bergsma, S., Post, M., and Yarowsky, D. Stylometric analysis of scientific articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2012), Association for Computational Linguistics, pp. 327–337.

[5] Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316* (2016).

[6] Coates, A., and Ng, A. Y. Learning feature representations with k-means. In *Neural networks: Tricks of the trade*. Springer, 2012, pp. 561–580.

[7] Cortes, C., and Vapnik, V. Support-vector networks. *Machine learning 20*, 3 (1995), 273–297.

[8] Cui, L., Huang, S., Wei, F., Tan, C., Duan, C., and Zhou, M. Superagent: A customer service chatbot for e-commerce websites. *Proceedings of ACL 2017, System Demonstrations* (2017), 97–102.

[9] Dave, K., Lawrence, S., and Pennock, D. M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web* (2003), ACM, pp. 519–528.

[10] de Vel, O. Y., Corney, M. W., Anderson, A. M., and Mohay, G. M. Language and gender author cohort analysis of e-mail for computer forensics. *Digital Forensic Research Conference* (2002).

[11] Durgesh, K. S., and Lekha, B. Data classification using support vector machine. *Journal of theoretical and applied information technology 12*, 1 (2010), 1–7.

[12] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. Liblinear: A library for large linear classification. *Journal of machine learning research 9*, Aug (2008), 1871–1874.

[13] Géron, A. *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems.* " O'Reilly Media, Inc.", 2017.

[14] Goldberg, A. B., Fillmore, N., Andrzejewski, D., Xu, Z., Gibson, B., and Zhu, X. May all your wishes come true: A study of wishes and how to recognize them. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (2009), Association for Computational Linguistics, pp. 263–271.

[15] Graves, A., Mohamed, A.-r., and Hinton, G. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing* (2013), IEEE, pp. 6645–6649.

[16] Hu, M., and Liu, B. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (2004), ACM, pp. 168–177.

[17] John, G. H., and Langley, P. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence* (1995), Morgan Kaufmann Publishers Inc., pp. 338–345.

[18] Kober, J., Bagnell, J. A., and Peters, J. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research 32*, 11 (2013), 1238–1274.

[19] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal 13* (2015), 8–17.

[20] Liu, D.-R., Lai, C.-H., and Lee, W.-J. A hybrid of sequential rules and collaborative filtering for product recommendation. *Information Sciences 179*, 20 (2009), 3505–3519.

[21] Liu, T.-Y., et al. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval 3*, 3 (2009), 225–331.

[22] Moghaddam, S. Beyond sentiment analysis: mining defects and improvements from customer feedback. In *European Conference on Information Retrieval* (2015), Springer, pp. 400–410.

[23] Murad, U., and Pinkas, G. Unsupervised profiling for identifying superimposed fraud. In *European Conference on Principles of Data Mining and Knowledge Discovery* (1999), Springer, pp. 251–261.

[24] Nguyen, T.-P., and Ho, T.-B. Detecting disease genes based on semi-supervised learning and protein–protein interaction networks. *Artificial intelligence in medicine 54*, 1 (2012), 63–71.

[25] Nigam, K., McCallum, A., Thrun, S., Mitchell, T., et al. Learning to classify text from labeled and unlabeled documents. *AAAI/IAAI 792* (1998), 6.

[26] Pang, B., Lee, L., et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval 2*, 1–2 (2008), 1–135.

[27] Press, S. J., and Wilson, S. Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association 73*, 364 (1978), 699–705.

[28] Ramanand, J., Bhavsar, K., and Pedanekar, N. Wishful thinking: finding suggestions and'buy'wishes from product reviews. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text* (2010), Association for Computational Linguistics, pp. 54–61.

[29] Tan, J., Kotov, A., Pir Mohammadiani, R., and Huo, Y. Sentence retrieval with sentiment-specific topical anchoring for review summarization. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (2017), ACM, pp. 2323–2326.

[30] Titov, I., and McDonald, R. A joint model of text and aspect ratings for sentiment summarization. *proceedings of ACL-08: HLT* (2008), 308–316.

[31] Yang, Z., Kotov, A., Mohan, A., and Lu, S. Parametric and non-parametric user-aware sentiment topic models. In *Proceedings of the 38th Inter-*

national ACM SIGIR Conference on Research and Development in Information Retrieval (2015), ACM, pp. 413–422.

[32] Yu, D., Varadarajan, B., Deng, L., and Acero, A. Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion. *Computer Speech & Language 24*, 3 (2010), 433–444.

[33] Yu, J., Zha, Z.-J., Wang, M., and Chua, T.-S. Aspect ranking: identifying important product aspects from online consumer reviews. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (2011), Association for Computational Linguistics, pp. 1496–1505.

[34] Zhou, Z., Li, X., and Zare, R. N. Optimizing chemical reactions with deep reinforcement learning. *ACS central science 3*, 12 (2017), 1337–1344.

[35] Zimek, A., Campello, R. J., and Sander, J. Ensembles for unsupervised outlier detection: challenges and research questions a position paper. *Acm Sigkdd Explorations Newsletter 15*, 1 (2014), 11–22.

ABSTRACT

METHODS FOR CLASSIFICATION OF CONSUMER REVIEW INTO THE ONES WRITTEN BEFORE OR AFTER THE PRODUCT PURCHASE

by

MD MEHEDI HASAN

May 2019

Advisor: Dr. Alexander Kotov

Major: Computer Science

Degree: Master of Science

Online consumer reviews provide a wealth of information about products and services that, if properly identified and extracted, could be of immense value to businesses. While classification of reviews according to sentiment polarity has been extensively studied in previous work, many more focused types of review analysis remain open problems. In this work, we introduce a novel text classification problem of separating post-purchase from pre-purchase consumer review fragments that can facilitate identification of immediate actionable insights based on the feedback from the customers, who actually purchased and own a product. To address this problem, we propose the features, which are based on the dictionaries and part-of-speech (POS) tags. Experimental results on the publicly available gold standard indicate that the proposed features allow to achieve nearly 75% accuracy for this problem and improve the performance of classifiers relative to using only lexical features.

AUTOBIOGRAPHICAL STATEMENT

MD MEHEDI HASAN

EDUCATION

- PhD Candidate (Computer Science)
  Wayne State University, Detroit, MI, USA
- Bachelor of Science (Computer Science and Engineering), 2009
  Bangladesh University of Engineering and Technology, Dhaka

PUBLICATIONS

1. **Hasan, M.**, Kotov, A., Naar, S., Alexander, G.L. and Carcone, A.I. "Deep neural architectures for discourse segmentation in email-based behavioral intervention". In Proceedings of the AMIA Informatics Summit (2019). American Medical Informatics Association.

2. Carcone, A.I., **Hasan, M.**, Alexander, G.L., Dong, M., Eggly, S., Brogan Hartlieb, K., Naar, S., MacDonell, K. and Kotov, A. "Developing machine learning models for behavioral coding". Journal of pediatric psychology (2019), 44(3), pp. 289–299.

3. **Hasan, M.**, Carcone, A.I., Naar, S., Eggly, S., Alexander, G., BroganHartlieb, K. and Kotov, A. "Identifying Effective Motivational Interviewing Communication Sequences Using Automated Pattern Analysis". Journal of Healthcare Informatics Research (JHIR) (2018), 3(1), pp. 86–106.

4. **Hasan, M.**, Kotov, A., Carcone, A.I., Dong, M. and Naar, S. "Predicting the Outcome of Patient-Provider Communication Sequences using Recurrent Neural Networks and Probabilistic Models". In Proceedings of the AMIA Informatics Summit (2018). American Medical Informatics Association, pp. 64–73.

5. **Hasan, M.**, Kotov, A., Carcone, A.I., Dong, M., Naar, S. and Hartlieb, K.B. "A study of the effectiveness of machine learning methods for classification of clinical interview fragments into a large number of categories". Journal of biomedical informatics (2016), 62, pp. 21–31.

6. **Hasan, M.**, Kotov, A., Mohan, A., Lu, S. and Stieg, P.M. "Feedback or Research: Separating Pre-purchase from Post-purchase Consumer Reviews". In European Conference on Information Retrieval (2016). Springer International Publishing, pp. 682–688.

7. Kotov, A., **Hasan, M.**, Carcone, A.I., Dong, M., Naar-King, S. and Brogan-Hartlieb, K. "Interpretable probabilistic latent variable models for automatic annotation of clinical text". In AMIA Annual Symposium Proceedings (2015). American Medical Informatics Association, pp. 785–794.