Sentiment Analysis on Twitter Based on Ensemble of Psychological and Linguistic Feature Sets

A. Onan

***Abstract*—With the advances in information and communication technologies, social media and microblogging platforms serve as an important source of information. In microblogging platforms, people can share their opinions, complaints, sentiments and attitudes towards topics, current issues and products. Sentiment analysis is an important research direction in natural language processing, which aims to identify the sentiment orientation of source materials. Twitter is a popular microblogging platform, where people all over the world can interact by user-generated text messages. Information obtained from Twitter can serve as an essential source for several applications, including event detection, news recommendation and crisis management. In sentiment classification, the identification of an appropriate feature subset plays an important role. LIWC (Linguistic Inquiry and Word Count) is an exploratory text analysis software to extract psycholinguistic features from text documents. In this paper, we present a psycholinguistic approach to sentiment analysis on Twitter. In this scheme, we utilized five main LIWC categories (namely, linguistic processes, psychological processes, personal concerns, spoken categories and punctuation) as feature sets. In the experimental analysis, five LIWC categories and their ensemble combinations are taken into consideration. To explore the predictive performance of different feature engineering schemes, four supervised learning algorithms (namely, Naïve Bayes, support vector machines, k-nearest neighbor algorithm and logistic regression) and three ensemble learning methods (namely, AdaBoost, Bagging and Random Subspace) are utilized. The experimental results indicate that ensemble feature sets yield higher predictive performance compared to the individual feature sets.**

Index Terms— **Machine learning, psychological feature sets, sentiment analysis, Twitter.**

I. INTRODUCTION

THE IMMENSE QUANTITY OF INFORMATION available with the remarkable growth of social media and microblogging platforms can serve as an essential source for decision making about products, services and policies [1-2].

Twitter is a popular and fast growing microblogging platform, where people can send short messages (referred as tweets) within a character limit of 140.

A. ONAN, is with Department of Software Engineering, Celal Bayar University, Manisa, Turkey, (e-mail: aytug.onan@cbu.edu.tr).
Manuscript received August 12, 2017; accepted Nov 16, 2017.
DOI: 10.17694/bajece. 419538

Twitter enables users to communicate in an efficient way. The user generated content on Twitter provide a useful source of information for researchers and practitioners [3]. Information obtained from Twitter can serve as an essential source of information for several applications, including event detection, epidemic dispersion, news recommendation and crisis management [4-6]. Sentiment analysis (also known as opinion mining) is an important research direction in natural language processing, which aims to identify the sentiment orientation of source materials. Sentiment analysis can be utilized for obtaining information regarding new products and services. It can be further applied to identify positive and negative aspects of a particular product or service [7].

The methods of sentiment analysis can be mainly divided into two groups as lexicon-based approaches and machine-learning based approaches. In addition, sentiment analysis can be conducted at different granularities based on the levels of details. Based on the levels of details, sentiment analysis methods are grouped into three categories as: document-level, sentence-level and aspect-level sentiment analysis [8].

Sentiment analysis can be modelled as a text classification problem. In machine learning based sentiment analysis, supervised classification algorithms (such as Naïve Bayes algorithm, support vector machines, k-nearest neighbor algorithm and logistic regression) can be utilized to identify sentiment orientation. Machine learning based sentiment analysis schemes involve data preprocessing, feature extraction and selection and training supervised classification algorithms with labelled data set.

In order to obtain a classification scheme with high predictive performance, feature extraction is an essential task [9]. LIWC (Linguistic Inquiry and Word Count) is an exploratory text analysis software to extract psycholinguistic features from text documents. Features related to psychological, linguistic, social and cultural aspects can be important for sentiment analysis [10]. For this purpose, we present a psycholinguistic approach to sentiment analysis on Twitter. In this paper, we utilized five main LIWC categories (namely, linguistic processes, psychological processes, personal concerns, spoken categories and punctuation) as feature sets. In the experimental analysis, five LIWC categories and their ensemble combinations are taken into consideration. To explore the predictive performance of

different feature engineering schemes, Naïve Bayes, support vector machines, k-nearest neighbor algorithm and logistic regression are utilized. In addition, ensemble methods (namely, Bagging, AdaBoost and Random Forest algorithms) are also considered to examine the predictive performance of supervised learning algorithms in conjunction with ensemble methods for sentiment analysis.

The rest of the paper is organized as follows: In Section 2, related work on sentiment analysis is presented. Section 3 presents the methodology of the study and Section 4 presents the experimental procedure and empirical results. Section 5 describes the concluding remarks.

II. Related Work

Sentiment analysis on Twitter data has attracted research attention. This section briefly reviews the existing works on sentiment analysis on Twitter data. Sentiment analysis on Twitter poses several challenges, due to the short length of messages and unstructured, informal and irregular nature of content. Hence, identification of an appropriate feature set is an important research direction. For instance, Go et al. [11] examined the usage of unigrams, bigrams, unigrams and bigrams and part of speech tags as features. In the classification phase, Naïve Bayes, maximum entropy and support vector machine classifiers were utilized. The empirical analysis indicated augmenting unigrams and bigrams yields better predictive performance compared to the other feature engineering schemes. Part of speech tags were not useful features and the highest classification performance was achieved by maximum entropy learner. In another study, Barbosa and Feng [12] explored the predictive performance of n-grams and tweet syntax features (such as retweets, hashtags, replies, links, punctuation, emoticons and upper cases). The empirical analysis on support vector machines indicated that tweet syntax features enhance the predictive performance of sentiment classification schemes on Twitter and n-grams cannot completely reveal the text messages. Similarly, Pak and Paroubek [13] examined the usage of n-grams and part of speech tags as features. In the empirical analysis, multinomial Naïve Bayes, support vector machines and conditional random field classifiers were utilized. The empirical analysis indicated that the utilization of part of speech tags in conjunction to n-grams yields better predictive performance on sentiment analysis of Twitter data. In another study, Koulumpis et al. [14] explored the usage of n-gram features, lexicon features, part of speech tags and microblogging features (such as the presence of positive, negative and neutral emoticons and the presence of intensifiers) on sentiment analysis of Twitter data. The experimental analysis indicated that the highest predictive performance among different feature engineering schemes was obtained by n-gram features in conjunction to lexicon features and microblogging features. In addition, the results indicated that integrating parts of speech features dropped the predictive performance of sentiment classification. Similarly, Agarwal et al. [15] examined the usage of part of speech features, lexicon features and microblogging features for sentiment analysis of Twitter data. In addition, they introduced a tree based representation to augment different feature engineering schemes in an efficient way. The experimental analysis indicated that the usage of prior polarity of words in conjunction with their part of speech tags yields the highest classification accuracy.

In another study, Saif et al. [16] examined the usage of unigram features, part of speech features and sentiment topic features for sentiment analysis on Twitter. The experimental analysis indicated that semantic feature set based approach yield better predictive performance compared to other feature engineering schemes.

Onan [1] examined the predictive performance of different n-gram models (namely, unigram, bigram and trigram) and their combinations on sentiment analysis of Turkish Twitter messages. In the empirical analysis, the highest predictive performance is achieved by the combination of unigram and bigram features. In another study, Salas-Zarate et al. [10] examined the performance of psycholinguistic feature sets in sentiment analysis of product reviews. In this study, linguistic processes, psychological processes, personal concerns, spoken categories and punctuation are taken into consideration. While existing work on sentiment analysis of Twitter data concentrates on n-grams, part of speech tags and microblogging based features, this study aims to examine the predictive performance of psychological and linguistic features obtained by LIWC on sentiment analysis on Twitter.

III. Methodology

This section describes dataset collection process, data processing, feature engineering schemes to represent the dataset, classification algorithms and ensemble learning methods utilized in the experimental analysis.

A. Dataset Collection

To evaluate the predictive performance of psychological and linguistic features on sentiment analysis, we have carried out an analysis on English messages on Twitter that contain positive, negative and neutral sentiments. In the dataset collection, we adopted the framework presented in [17]. We utilized Twitter4J, an open-source Java library for utilizing Twitter Streaming API, to collect tweets. Each tweet is labelled by a single class label, either as positive, negative or neutral. After collecting the tweets, automatic filtering was applied to remove irrelevant and redundant tweets (retweets and duplicates). In this way, we obtained a collection of 6218 negative, 4891 positive and 4252 neutral tweets. In order to obtain a balanced corpus, our final dataset contains a collection of 4200 negative, 4200 positive and 4200 neutral tweets.

B. Data Preprocessing

Due to irregular and informal nature of Twitter messages, it is essential to preprocess the tweets so that particular problems (such as initialisms, unnecessary repetitions and misuse of letters) can be eliminated [18]. In the preprocessing stage, we adopted the framework presented in [19]. The preprocessing stage mainly seeks to remove unnecessary characters or

sequences, which have no value to the sentiment classification. For this purpose, the following tasks were performed on each tweet [19]:

- Remove mentions and replies to other users' tweets, which are represented by strings starting with "@".
- Remove URLs (namely, strings staring with "http://").
- Remove "#" character.

C. Feature Engineering

In this section, we examine different psycholinguistic feature sets on sentiment analysis. In this scheme, we utilized LIWC (Linguistic Inquiry and Word Count) to extract psycholinguistic features from the dataset. LIWC categories have been successfully utilized in several fields of computational linguistics, including sarcasm identification and satire detection [20, 21].

LIWC is a text analysis application to identify emotional, cognitive and structural aspects of verbal and written speech samples. The first version of LIWC application was developed in 1993 and the most recent version was released on 2015 [22]. LIWC contains dictionaries on several languages, including English, Spanish, German, Dutch, Norwegian, Italian and Portuguese.

LIWC Dictionary contains approximately 6400 words, word stems and emoticons. Each entry of the dictionary contains one or several word categories or sub-dictionaries. For a particular word encountered in the text, scores for the corresponding categories or dictionaries are incremented. The categories can be further classified into five main sets as linguistic processes, psychological processes, personal concerns, spoken categories and punctuation. In Table 1, main LIWC sets and categories are listed.

TABLE I
MAIN LIWC SETS AND CATEGORIES

Feature Set	Categories
Linguistic Processes	Word count, total pronouns, personal pronouns, articles, prepositions, auxiliary verbs, adverbs, conjunctions
Psychological Processes	Affective processes, positive emotion, negative emotion, social processes, cognitive processes, perceptual processes
Personal Concerns	Work, leisure, home, money
Spoken Categories	Assent, Non-fluencies, fillers
Punctuation	Total punctuation, periods, commas, colons, semicolons, question marks, exclamation marks, dashes

As it can be observed from the categories listed in Table 1, linguistic processes contains grammatical information, such as word count, total number of pronouns, personal pronouns, articles, prepositions and auxiliary verbs. Psychological processes involves psychological information, such as affective processes, positive emotion, and negative emotion and so on. Personal concerns contains information, such as work, leisure, home and money. Spoken categories involves information regarding the spoken language. Finally, punctuation set involves punctuation marks, such as punctuation, periods, commas, colons, semicolons, question marks, exclamation marks, dashes.

Based on the aforementioned five main LIWC feature sets, target words or word stems are searched through the LIWC dictionary. Each word is assigned to one or more sub-dictionaries.

D. Classification Algorithms

To evaluate the predictive performance of different feature engineering schemes, Naïve Bayes, support vector machines, k-nearest neighbor algorithm and logistic regression algorithm are utilized.

Naïve Bayes algorithm (NB) is a probabilistic classification algorithm based on Bayes' theorem. It has a simple structure due to the assumption of conditional independence. Though its simple structure, it can be effectively utilized in text and web mining applications [23].

Support vector machines (SVM) are supervised learning algorithms that can be utilized to solve classification and regression problems. They can be applied effectively to classify both linear and non-linear data [24]. Support vector machines build a hyperplane in a higher dimensional space to solve classification or regression problem. The hyperplane aims to make a good separation by achieving the largest distance to the nearest training data points of classes (known as functional margin).

K-nearest neighbor algorithm (KNN) is an instance-based classifier. In KNN algorithm, the class label of each instance is determined based on the k-nearest neighbors of the instance. Based on the predictions of the neighbor instances, a majority voting scheme is utilized to determine the class label.

Logistic regression (LR) is a linear classification algorithm, which uses a linear function of a set of predictor variables to model the probability of some event's occurring [25]. Linear regression can yield good results. However, the membership values generated by linear regression cannot be always in [0-1] range, which is not an appropriate range for probabilities. In logistic regression, a linear model is constructed on the transformed target variable whilst eliminating the mentioned problems.

E. Ensemble Learning Methods

This section briefly describes the ensemble learning algorithms utilized in the empirical analysis.

Bagging (Bootstrap aggregating) is a popular ensemble learning method, which aims to obtain a single prediction with higher predictive performance by combining weak learning algorithms trained on different training sets [26]. In this scheme, different training sets are obtained by simple random sampling with replacement. The predictions of weak learning algorithms are combined by majority voting or weighted voting.

AdaBoost algorithm is another popular ensemble learning method, which aims to obtain a robust classification scheme by focusing on the data points that are difficult to classify [27]. In this scheme, the weight values assigned to the instances of the training set are adjusted so that the weight values of misclassified instances are increased, whereas the

weight values of correctly classified instances are decreased. In this way, the learning algorithms focus on classifying the difficult instances.

Random subspace algorithm is an ensemble learning algorithm which combines multiple classifiers trained on the randomly selected feature subspaces [28]. The algorithm aims to avoid over-fitting, while providing high predictive performance by training the weak learning algorithms on different samples of the feature space.

IV. Experimental Analysis And Results

This section presents the evaluation measures, experimental procedure and the experimental results of the study.

A. Evaluation Measures

In order to evaluate the performance of classification algorithms, two different evaluation measures, namely, classification accuracy and F-measure.

Classification accuracy (ACC) is the proportion of true positives and true negatives obtained by the classification algorithm over the total number of instances as given by Equation 1 [29]:

$$ACC = \frac{TN + TP}{TP + FP + FN + TN} \quad (1)$$

where TN denotes number of true negatives, TP denotes number of true positives, FP denotes number of false positives and FN denotes number of false negatives.

Precision (PRE) is the proportion of the true positives against the true positives and false positives as given by Equation 2:

$$PRE = \frac{TP}{TP + FP} \quad (2)$$

Recall (REC) is the proportion of the true positives against the true positives and false negatives as given by Equation 3:

$$REC = \frac{TP}{TP + FN} \quad (3)$$

F-measure takes values between 0 and 1. It is the harmonic mean of precision and recall as determined by Equation 4:

$$F - measure = \frac{2 * PRE * REC}{PRE + REC} \quad (4)$$

B. Experimental Procedure

In the experimental analysis, 10-fold cross validation method is employed. In this scheme, the original dataset is randomly divided into ten mutually exclusive folds. Training and testing process is repeated ten times and each part is tested and trained ten times and the average results for 10-fold are reported. The experimental analysis is performed with the machine learning toolkit WEKA (Waikato Environment for Knowledge Analysis) version 3.9, which is an open-source platform that contains many machine learning algorithms implemented in JAVA.

C. Experimental Results

In Tables 2-3, classification accuracies and F-measure results obtained by psycholinguistic feature sets and the four base learning algorithms are presented, respectively.

TABLE II
CLASSIFICATION RESULTS OBTAINED BY SUPERVISED LEARNING METHODS ON PSYCHOLINGUISTIC FEATURE SETS

Feature set	NB	SVM	KNN	LR
LP	77.35	76.73	72.30	72.80
PP	77.28	76.57	72.17	72.06
PC	76.40	75.77	71.97	71.76
SC	74.57	75.66	70.88	71.55
PU	74.25	75.60	70.56	71.54
LP+PP	79.41	78.06	73.86	76.52
LP+PC	79.35	78.06	73.85	76.43
LP+SC	79.25	78.04	73.84	76.32
LP+PU	79.17	77.95	73.80	76.06
PP+PC	79.14	77.92	73.51	76.05
PP+SC	77.98	77.33	72.86	74.62
PP+PU	77.71	77.10	72.82	74.59
PC+SC	77.63	76.98	72.78	74.55
PC+PU	77.50	76.91	72.78	73.95
SC+PU	77.36	76.82	72.50	72.99
LP+PP+PC	**86.94**	**82.50**	**79.39**	**81.09**
LP+PP+SC	80.49	79.12	75.70	78.04
LP+PP+PU	80.44	79.10	75.53	78.01
LP+PC+SC	80.29	79.05	75.45	77.71
LP+PC+PU	80.24	78.94	75.28	77.69
LP+SC+PU	80.04	78.81	75.09	77.56
PP+PC+SC	79.90	78.61	74.99	77.52
PP+PC+PU	79.81	78.55	74.78	77.46
PP+SC+PU	78.75	77.82	73.27	75.92
PC+SC+PU	78.69	77.60	73.25	75.85
LP+PP+PC+SC	85.24	82.02	78.33	80.81
LP+PP+PC+PU	84.56	81.96	77.98	80.70
LP+PP+SC+PU	83.88	81.75	77.59	80.61
LP+PC+SC+PU	83.81	81.55	77.27	80.09
PP+PC+SC+PU	83.38	81.34	77.24	80.01
LP+PP+PC+SC+PU	83.20	81.08	77.14	79.27

NB: Naïve Bayes algorithm, SVM: support vector machines, KNN: K-nearest neighbor algorithm, LR: logistic regression, LP: linguistic processes, PP: psychological processes, PC: personal concerns, SC: Spoken categories, PU: punctuation.

In the first column, the different dimensions of LIWC used for training a particular classifier are reported. For instance, LP+PP indicates that linguistic processes and psychological processes are taken into account in the empirical analysis.

TABLE III
F-MEASURE VALUES OBTAINED BY SUPERVISED LEARNING METHODS ON PSYCHOLINGUISTIC FEATURE SETS

Feature set	NB	SVM	KNN	LR
LP	0.78	0.77	0.73	0.73
PP	0.78	0.77	0.72	0.72
PC	0.77	0.76	0.72	0.72
SC	0.75	0.76	0.71	0.72
PU	0.75	0.76	0.71	0.72
LP+PP	0.80	0.78	0.74	0.77
LP+PC	0.80	0.78	0.74	0.77
LP+SC	0.80	0.78	0.74	0.77
LP+PU	0.79	0.78	0.74	0.76
PP+PC	0.79	0.78	0.74	0.76
PP+SC	0.78	0.78	0.73	0.75
PP+PU	0.78	0.77	0.73	0.75
PC+SC	0.78	0.77	0.73	0.75
PC+PU	0.78	0.77	0.73	0.74
SC+PU	0.78	0.77	0.73	0.73
LP+PP+PC	**0.87**	**0.83**	**0.80**	**0.81**
LP+PP+SC	0.81	0.79	0.76	0.78
LP+PP+PU	0.81	0.79	0.76	0.78
LP+PC+SC	0.81	0.79	0.76	0.78
LP+PC+PU	0.81	0.79	0.76	0.78
LP+SC+PU	0.80	0.79	0.75	0.78
PP+PC+SC	0.80	0.79	0.75	0.78
PP+PC+PU	0.80	0.79	0.75	0.78
PP+SC+PU	0.79	0.78	0.74	0.76
PC+SC+PU	0.79	0.78	0.74	0.76
LP+PP+PC+SC	0.86	0.82	0.79	0.81
LP+PP+PC+PU	0.85	0.82	0.78	0.81
LP+PP+SC+PU	0.84	0.82	0.78	0.81
LP+PC+SC+PU	0.84	0.82	0.78	0.80
PP+PC+SC+PU	0.84	0.82	0.78	0.80
LP+PP+PC+SC+PU	0.84	0.81	0.77	0.80

NB: Naïve Bayes algorithm, SVM: support vector machines, KNN: K-nearest neighbor algorithm, LR: logistic regression, LP: linguistic processes, PP: psychological processes, PC: personal concerns, SC: Spoken categories, PU: punctuation.

Considering the experimental results (in terms of classification accuracies and F-measure values) presented in Tables 2-3, different classification algorithms have different predictive performance. The highest predictive performance among the compared classifiers is achieved by Naïve Bayes algorithm and the second highest predictive performance is obtained by support vector machines. Logistic regression classifier and K-nearest neighbour algorithm generally yield similar predictive performance.

The study seeks to examine the predictive performance of different LIWC categories (namely, linguistic processes, psychological processes, personal concerns, spoken categories and punctuation) and their subsets as feature sets.

Regarding the predictive performance of individual feature sets, the highest predictive performance is achieved by using linguistic processes (denoted as LP). The second highest predictive performance is obtained by psychological processes, the third highest predictive performance is obtained by personal concerns and the lowest predictive performance is obtained by punctuation. Hence, linguistic processes, psychological processes and personal concerns provide clues to better analysis sentiment on Twitter. The highest predictive performance (77.35%) by the individual feature sets is achieved by linguistic processes and Naïve Bayes algorithm.

As it can be observed from the results listed in Tables 2-3, ensemble feature sets (combining different LIWC dimensions) yield higher predictive performance compared to the individual feature sets. The highest predictive performance among the ensemble feature sets is obtained by combining linguistic processes, psychological processes and personal concerns. The highest predictive performance achieved by this configuration is 86.94%, it is utilized in conjunction to Naïve Bayes classifier.

Ensemble learning methods can be utilized to further enhance the predictive performance of supervised learning algorithms. In the empirical analysis, we have also considered ensemble of supervised learning algorithms in conjunction with psycholinguistic feature sets. In this regard, twelve ensemble schemes (AdaBoost, Bagging and Random Subspace ensembles of four supervised learning algorithms) are considered.

In Table 4, classification accuracies obtained by ensembles of feature sets and classifiers are presented. As it can be observed from the results listed in Table 4, the predictive performances of supervised learning algorithms are generally improved by using ensemble learning methods. Regarding the predictive performance of different ensemble learning methods, Random Subspace method generally yield better results than other ensemble learning methods. The highest predictive performance on the ensemble learning methods is achieved by the ensemble feature sets combining linguistic processes, psychological processes and personal concerns. For this configuration, Random Subspace ensemble of Naïve Bayes ensemble is utilized as the classifier. This configuration achieves a classification accuracy of 89.10%.

In Table 5, F-measure values obtained by ensembles of feature sets and classifiers are presented. The predictive performance patterns obtained in terms of classification accuracies are also valid for F-measure values listed in Table 5.

To summarize the main findings of the study, Figure 1 and Figure 2 depict the main effect plot for classification accuracy and the main effect plot for F-measure values, respectively.

TABLE IV
CLASSIFICATION RESULTS OBTAINED BY ENSEMBLE LEARNING METHODS ON PSYCHOLINGUISTIC FEATURE SETS

Ensemble Method	Bagging	Random Subspace	Ada-Boost	Bagging	Random Subspace	Ada-Boost	Bagging	Random Subspace	Ada-Boost	Bagging	Random Subspace	Ada-Boost
Base Learner	**NB**	**NB**	**NB**	**SVM**	**SVM**	**SVM**	**KNN**	**KNN**	**KNN**	**LR**	**LR**	**LR**
LP	78.10	79.51	78.20	77.46	77.63	77.49	73.11	73.13	73.08	73.64	73.71	73.64
PP	78.04	79.42	78.16	77.25	77.48	77.35	72.99	73.00	72.98	72.91	73.01	72.93
PC	77.14	78.54	77.29	76.47	76.69	76.58	72.82	72.78	72.76	72.59	72.67	72.67
SC	75.34	76.74	75.48	76.40	76.53	76.49	71.71	71.71	71.70	72.40	72.47	72.44
PU	75.00	76.40	75.18	76.26	76.47	76.39	71.39	71.39	71.37	72.38	72.43	72.42
LP+PP	80.16	81.59	80.31	78.79	78.97	78.87	74.64	74.71	74.65	77.41	77.41	77.40
LP+PC	80.11	81.53	80.25	78.76	78.97	78.85	74.64	74.66	74.65	77.26	77.32	77.30
LP+SC	80.02	81.44	80.15	78.75	78.96	78.84	74.64	74.67	74.61	77.19	77.22	77.19
LP+PU	79.92	81.35	80.15	78.65	78.88	78.75	74.65	74.68	74.60	76.91	77.00	76.94
PP+PC	79.91	81.31	80.01	78.60	78.81	78.75	74.32	74.34	74.29	76.89	76.98	76.95
PP+SC	78.76	80.17	78.88	78.05	78.25	78.12	73.68	73.70	73.67	75.49	75.51	75.48
PP+PU	78.46	79.91	78.60	77.77	78.00	77.90	73.61	73.66	73.65	75.42	75.52	75.47
PC+SC	78.34	79.81	78.54	77.71	77.87	77.76	73.59	73.64	73.59	75.40	75.46	75.41
PC+PU	78.27	79.67	78.44	77.64	77.79	77.71	73.63	73.64	73.55	74.83	74.87	74.80
SC+PU	78.10	79.52	78.22	77.56	77.75	77.59	73.35	73.33	73.28	73.82	73.90	73.88
LP+PP+PC	**87.68**	**89.10**	**87.82**	**83.19**	**83.38**	**83.33**	**80.19**	**80.22**	**80.17**	**81.92**	**82.03**	**81.94**
LP+PP+SC	81.25	82.68	81.40	79.80	79.98	79.94	76.53	76.50	76.52	78.86	78.92	78.91
LP+PP+PU	81.18	82.61	81.37	79.78	79.95	79.90	76.33	76.34	76.33	78.86	78.93	78.89
LP+PC+SC	81.03	82.42	81.24	79.69	79.95	79.85	76.28	76.29	76.26	78.57	78.62	78.59
LP+PC+PU	80.99	82.45	81.13	79.65	79.82	79.68	76.11	76.13	76.07	78.54	78.57	78.55
LP+SC+PU	80.82	82.25	80.96	79.57	79.72	79.56	75.93	75.90	75.88	78.40	78.47	78.43
PP+PC+SC	80.65	82.07	80.80	79.34	79.49	79.41	75.83	75.81	75.83	78.36	78.47	78.40
PP+PC+PU	80.56	81.98	80.67	79.26	79.48	79.37	75.58	75.62	75.59	78.27	78.34	78.35
PP+SC+PU	79.51	80.97	79.65	78.56	78.69	78.60	74.10	74.11	74.05	76.77	76.81	76.76
PC+SC+PU	79.43	80.90	79.62	78.32	78.50	78.39	74.08	74.08	74.03	76.70	76.77	76.74
LP+PP+PC+SC	85.99	87.41	86.12	82.72	82.95	82.79	79.14	79.15	79.13	81.65	81.73	81.71
LP+PP+PC+PU	85.32	86.74	85.46	82.61	82.87	82.79	78.78	78.81	78.83	81.54	81.63	81.53
LP+PP+SC+PU	84.61	86.02	84.77	82.48	82.67	82.58	78.44	78.39	78.44	81.46	81.53	81.44
LP+PC+SC+PU	84.55	85.96	84.73	82.22	82.44	82.34	78.09	78.06	78.06	80.97	81.03	80.96
PP+PC+SC+PU	84.11	85.56	84.32	82.02	82.23	82.15	78.06	78.06	78.01	80.90	80.93	80.86
LP+PP+PC+SC+PU	83.94	85.39	84.06	81.80	82.02	81.89	77.96	77.99	77.99	80.13	80.20	80.15

TABLE V
F-MEASURE VALUES OBTAINED BY ENSEMBLE LEARNING METHODS ON PSYCHOLINGUISTIC FEATURE SETS

Ensemble Method	Bagging	Random Subspace	Ada-Boost	Bagging	Random Subspace	Ada-Boost	Bagging	Random Subspace	Ada-Boost	Bagging	Random Subspace	Ada-Boost
Base Learner	**NB**	**NB**	**NB**	**SVM**	**SVM**	**SVM**	**KNN**	**KNN**	**KNN**	**LR**	**LR**	**LR**
LP	0.80	0.81	0.80	0.79	0.79	0.79	0.75	0.75	0.75	0.75	0.75	0.75
PP	0.80	0.81	0.80	0.79	0.79	0.79	0.74	0.74	0.74	0.74	0.74	0.74
PC	0.79	0.80	0.79	0.78	0.78	0.78	0.74	0.74	0.74	0.74	0.74	0.74
SC	0.77	0.78	0.77	0.78	0.78	0.78	0.73	0.73	0.73	0.74	0.74	0.74
PU	0.77	0.78	0.77	0.78	0.78	0.78	0.73	0.73	0.73	0.74	0.74	0.74
LP+PP	0.82	0.83	0.82	0.80	0.81	0.80	0.76	0.76	0.76	0.79	0.79	0.79
LP+PC	0.82	0.83	0.82	0.80	0.81	0.80	0.76	0.76	0.76	0.79	0.79	0.79
LP+SC	0.82	0.83	0.82	0.80	0.81	0.80	0.76	0.76	0.76	0.79	0.79	0.79
LP+PU	0.82	0.83	0.82	0.80	0.80	0.80	0.76	0.76	0.76	0.78	0.79	0.79
PP+PC	0.82	0.83	0.82	0.80	0.80	0.80	0.76	0.76	0.76	0.78	0.79	0.79
PP+SC	0.80	0.82	0.80	0.80	0.80	0.80	0.75	0.75	0.75	0.77	0.77	0.77
PP+PU	0.80	0.82	0.80	0.79	0.80	0.79	0.75	0.75	0.75	0.77	0.77	0.77
PC+SC	0.80	0.81	0.80	0.79	0.79	0.79	0.75	0.75	0.75	0.77	0.77	0.77
PC+PU	0.80	0.81	0.80	0.79	0.79	0.79	0.75	0.75	0.75	0.76	0.76	0.76
SC+PU	0.80	0.81	0.80	0.79	0.79	0.79	0.75	0.75	0.75	0.75	0.75	0.75
LP+PP+PC	**0.89**	**0.91**	**0.90**	**0.85**	**0.85**	**0.85**	**0.82**	**0.82**	**0.82**	**0.84**	**0.84**	**0.84**
LP+PP+SC	0.83	0.84	0.83	0.81	0.82	0.82	0.78	0.78	0.78	0.80	0.81	0.81
LP+PP+PU	0.83	0.84	0.83	0.81	0.82	0.82	0.78	0.78	0.78	0.80	0.81	0.81
LP+PC+SC	0.83	0.84	0.83	0.81	0.82	0.81	0.78	0.78	0.78	0.80	0.80	0.80
LP+PC+PU	0.83	0.84	0.83	0.81	0.81	0.81	0.78	0.78	0.78	0.80	0.80	0.80
LP+SC+PU	0.82	0.84	0.83	0.81	0.81	0.81	0.77	0.77	0.77	0.80	0.80	0.80
PP+PC+SC	0.82	0.84	0.82	0.81	0.81	0.81	0.77	0.77	0.77	0.80	0.80	0.80
PP+PC+PU	0.82	0.84	0.82	0.81	0.81	0.81	0.77	0.77	0.77	0.80	0.80	0.80
PP+SC+PU	0.81	0.83	0.81	0.80	0.80	0.80	0.76	0.76	0.76	0.78	0.78	0.78
PC+SC+PU	0.81	0.83	0.81	0.80	0.80	0.80	0.76	0.76	0.76	0.78	0.78	0.78
LP+PP+PC+SC	0.88	0.89	0.88	0.84	0.85	0.84	0.81	0.81	0.81	0.83	0.83	0.83
LP+PP+PC+PU	0.87	0.89	0.87	0.84	0.85	0.84	0.80	0.80	0.80	0.83	0.83	0.83
LP+PP+SC+PU	0.86	0.88	0.87	0.84	0.84	0.84	0.80	0.80	0.80	0.83	0.83	0.83
LP+PC+SC+PU	0.86	0.88	0.86	0.84	0.84	0.84	0.80	0.80	0.80	0.83	0.83	0.83
PP+PC+SC+PU	0.86	0.87	0.86	0.84	0.84	0.84	0.80	0.80	0.80	0.83	0.83	0.83
LP+PP+PC+SC+PU	0.86	0.87	0.86	0.83	0.84	0.84	0.80	0.80	0.80	0.82	0.82	0.82

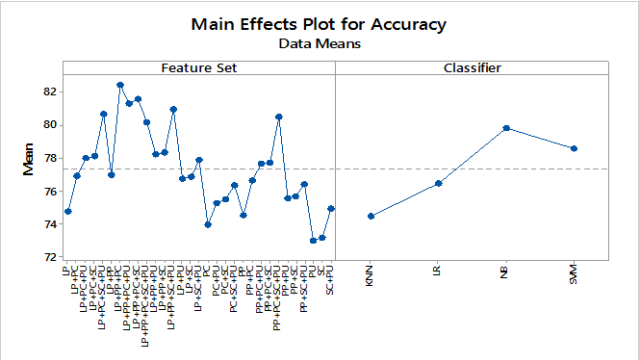

Fig.1. The main effects plot for accuracy

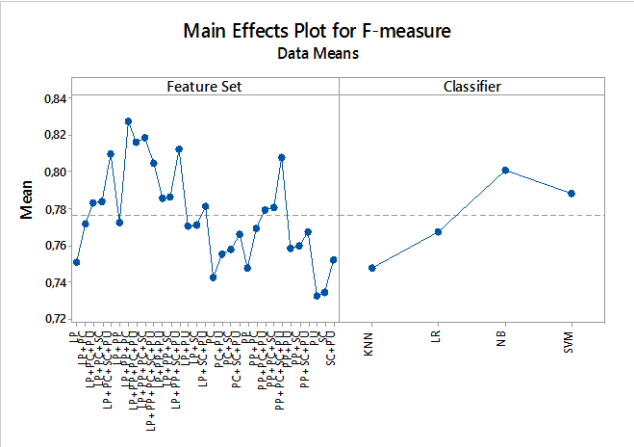

Fig.2. The main effects plot for F-measure

V. CONCLUSION

Social media and microblogging platforms serve as an essential source of information. Sentiment analysis on Twitter is a promising research direction. Sentiment analysis on Twitter is a challenging problem, where unstructured, informal and irregular content should be properly handled. The identification of an appropriate feature set is important to build classification schemes with high predictive performance. In the earlier work on sentiment analysis of Twitter data, n-grams, part of speech tags and microblogging based features are considered.

In this paper, we examined the predictive performance of psychological and linguistic features obtained by LIWC on sentiment analysis on Twitter. For this purpose, five main LIWC categories (namely, linguistic processes, psychological processes, personal concerns, spoken categories and punctuation) and their combinations are taken as feature sets. The experimental analysis with classification algorithms indicate that psycholinguistic feature sets can yield encouraging results on sentiment analysis of Twitter data. The experimental analysis indicates that ensemble feature sets outperforms the individual feature sets. For sentiment analysis on Twitter, the highest predictive performance (89.10%) is achieved by by combining linguistic processes, psychological processes and personal concerns with Random Subspace ensemble of Naïve Bayes.

REFERENCES

[1] A. Onan, "Twitter mesajları üzerinde makine öğrenmesi yöntemlerine dayalı duygu analizi", Yönetim Bilişim Sistemleri Dergisi, Vol. 3, No. 2, 2017, pp. 1-14.

[2] A. Onan, S. Korukoğlu, and H. Bulut, "A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification", Expert Systems with Applications, Vol.62, 2016, pp.1-16.

[3] A.Onan, "A machine learning based approach to identify geo-location of Twitter users", in *Proceedings of the ICC 2017*, UK, 2017, pp.1-7.

[4] J. Mahmud, J. Nichols, and C. Drews, "Home location identification of twitter users", ACM Transactions on Intelligent Systems and Technology, Vol. 5, No.3, 2014, pp.47.

[5] Z. Cheng, J. Caverlee, and K.Lee, "You are where you tweet: a content-based approach to geo-location twitter users", in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, USA, 2010, pp.759-768.

[6] B.Hecht, L.Hong, B. Suh and E.D.Chi, "Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles", in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, USA, 2011, pp.237-246.

[7] A. Onan and S. Korukoğlu, "Makine öğrenmesi yöntemlerinin görüş madenciliğinde kullanılması üzerine bir literatür araştırması", Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, Vol. 22, No. 2, 2016, pp. 111-122.

[8] W. Medhat, A. Hassan and H. Korashy, "Sentiment analysis algorithms and applications: a survey", Ain Shams Engineering Journal, Vol. 5, No. 4, 2014, pp. 1093-1113.

[9] A. Onan and S. Korukoğlu, "A feature selection model based on genetic rank aggregation for text sentiment classification", Journal of Information Science, Vol. 43, No.1, 2017, pp.25-38.

[10] M.P. Salas-Zarate, E.Lopez-Lopez, R.Valencia-Garcia, N. Gilles, A.Almela and G.Alor-Hernandez, "A study on LIWC categories for opinion mining in Spanish reviews", Journal of Information Science, Vol.40, No.6, 2014, pp.749-760.

[11] A.Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision", *CS224N Project Report*, 2009.

[12] L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data", in *Proceedings of ACL*, USA, 2010, pp. 36-44.

[13] A.Pak and P.Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining", in *Proceedings of LREC 2010*, USA, 2010, pp. 1320-1326.

[14] E. Kouloumpis, T.Wilson and J.D.Moore, "Twitter sentiment analysis: the good, the bad and the omg!", in *Proceedings of ICWSM 2011*, USA, 2011, pp. 538-541.

[15] A.Agarwal, B.Xie, I.Vovsha, O.Rambow and R. Passonneau, "Sentiment analysis of twitter data", in *Proceedings of ACL 2011*, USA, 2011, pp. 30-38.

[16] H.Saif, Y.He and H.Alani, "Semantic sentiment analysis of twitter", in *Proceedings of ISWC 2012*, USA, 2012, pp.508-524.

[17] M.Salas-Zarate, M.A. Paredes-Valverde, M.A.Rodriguez-Garcia, R.Valencia-Garcia and G.Alor-Hernandez, "Automatic detection of satire in Twitter: a psycholinguistic-based approach", Knowledge-Based Systems, Vol.128, 2017, pp.20-33.

[18] J.M.Cotelo, F.L.Cruz, J.A.Troyano and F.J.Ortega, "A modular approach for lexical normalization applied to Spanish tweets", Expert Systems with Applications, Vol. 42, No.10, 2015,pp. 4743-4754.

[19] E.Kontopoulos, C.Berberidis, T.Dergiades and N.Bassiliades, "Ontolog-based sentiment analysis of twitter posts", Expert Systems with Applications, Vol.40, No.10, 2013, pp.4065-4074.

[20] R.Justo, T.Corcoran, S.M.Lukin, M.Walker and M.I.Torres, "Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web", Knowledge-Based Systems, Vol. 69, 2014, pp.124-133.

[21] S.Skalicky and S.Crossley, "A statistical analysis of satirical Amazon.com product reviews", European Journal of Humour Research, Vol.2, 2015, pp.66-85.

[22] J.W.Pennebaker, R.L.Boyd, K.Jordan and K.Blackburn, "The development and psychometric properties of LIWC 2015".

[23] A.Onan, "Classifier and feature set ensembles for web page classification", Journal of Information Science, Vol. 42, No.2, pp.150-165.

[24] A.Onan, "Sarcasm identification on twitter: a machine learning approach", in *Proceedings of CSOC 2017*, Germany, 2017, pp.374-383.

[25] M.Kantardzic, *Data mining: concepts, models, methods and algorithms*, John Wiley & Sons, 2011, p.552.

[26] L.Breiman, "Bagging predictors", Machine Learning, Vol.4, No.2, pp.123-140.

[27] Y.Freund and R.E.Schapire, "Experiments with a new boosting algorithm", in *Proceedings of the Thirteenth International Conference on Machine Learning*, Italy, 1996, pp.148-156.

[28] T.K. Ho, "The random subspace method for constructing decision forests", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No.8, pp.832-844.

[29] A.Onan, "Artificial immune system based web page classification", in *Proceedings of CSOC 2015*, Germany, 2015, pp.189-199.

BIOGRAPHIES

AYTUĞ ONAN was born in 1987. Dr. Aytuğ Onan received his BS. in Computer Engineering from Izmir University of Economics (Turkey) in 2010. He earned his MS in Computer Engineering and PhD in Computer Engineering from Ege University (Turkey) in 2013 and 2016, respectively. He has been working as "assistant professor" since January 2017 at the Department of Software Engineering of Celal Bayar University (Turkey). He has been reviewing for several international journals, including Expert Systems with Applications, Plos One, International Journal of Machine Learning and Cybernetics, Journal of Information Science. He has published several journal papers on machine learning and computational linguistics.