# A Novel Classifier Based on Meaning for Text Classification

Murat Can Ganiz[1,2], Melike Tutkan[1], Selim Akyokuş[1]
[1] Computer Engineering Department of
Doğuş University, Computer Engineering Dept., İstanbul, Turkey
[2] VeriUs Teknoloji, İstanbul, Turkey
{mcganiz,mtutkan,sakyokus}@dogus.edu.tr

*Abstract*— **Text classification is one of the key methods used in text mining. Generally, traditional classification algorithms from machine learning field are used in text classification. These algorithms are primarily designed for structured data. In this paper, we propose a new classifier for textual data, called Supervised Meaning Classifier (SMC). The new SMC classifier uses meaning measure, which is based on Helmholtz principle from Gestalt Theory. In SMC, meaningfulness of terms in the context of classes are calculated and used for classification of a document. Experiment results show that new SMC classifier outperforms traditional classifiers of Multinomial Naïve Bayes (MNB) and Support Vector Machine (SVM) especially when the training data limited.**

***Keywords—machine learning; Helmholtz Principle; text classification; sentiment analysis.***

## I. INTRODUCTION

Text classification (or text categorization) is the process of assigning class labels or categories from a predefined set to the natural language documents according to their content. In recent years, there has been an explosive growth of textual data on the Internet and organizations, and the volume of textual data continues to increase every day due to extensive usage of information and Internet technologies. Due to such large and increasing volumes of textual data, text classification is attracting interest of researchers in many fields. In the literature, there are many machine learning algorithms proposed for classification such as Support Vector Machine (SVM) [2], Naive Bayes (NB) [3], K-Nearest Neighbor (k-NN) [4], Decision Trees (DT) [5], Multinomial Naïve Bayes (MNB) [6] and theirs variants. In a study presented in the IEEE International Conference on Data Mining (ICDM) in December 2006 [22], the top ten influential data mining algorithms include the following classification algorithms: C4.5 (DT), SVM, k-NN, Naive Bayes, and CART. Among them SVM and MNB are by far the most commonly used ones in text classification domain. More information about text classification algorithms can be found in [23] and also in [24].

In this study, we propose a novel text classifier, called Supervised Meaning Classifier (SMC). It uses meaning measure, which is based on Helmholtz Principle from Gestalt Theory of human perception [11]. In brief, Helmholtz Principle from Gestalt Theory claims if an observed geometric structure has very low probability to appear in noise, this geometric structure is perceptually meaningful; this means that if an unexpected event happens in a particular context, humans can easily notice it. By using Helmholtz Principle from Gestalt Theory, Balinsky et al. [15] presents and defines the "meaning measure" to be used in textual, unstructured and sequential data mining applications. This measure uses the fact that interesting events appear as large deviations from randomness.

Helmholtz Principle from Gestalt theory based meaning measure is previously used in unusual behavior detection and information extraction from small documents [13], in automatic text summarization [16] by defining relations between sentences using social network analysis and properties of small world phenomenon [17], in rapid change detection in data streams and documents [14], in keyword extraction and rapid change detection [15], in extractive text summarization by modeling texts and documents as a small world networks [18] and in automatic text and data stream segmentation [12].

In our previous studies, we adopted and applied meaning measure for supervised and unsupervised feature selection [8][9][10] and as a semantic kernel [25]. In this study we propose a new classifier, which uses the supervised variant of the meaning measure which is proposed in [8] and [9]. In SMC, meaningfulness of terms in the context of classes are calculated and used for classification of a document.

The remainder of this paper is organized as follows: Section 2 presents some information about well-known classifiers and introduces Helmholtz principle, Section 3 presents and analyzes the proposed classifier algorithm, Section 4 presents experimental setup and introduces datasets, Section 5 presents experiment results including some opinions and the last section presents a conclusion and future work.

## II. RELATED WORK

### A. Classification

There are various classifiers used for text classification. The most popular one is Support Vector Machines (SVM), which is a discriminative and binary classifier that finds an optimal hyperplane by maximizing the margins among the closest points of instances in different classes. It employs an optimization method called as quadratic programming. Another popular

classifier is k-Nearest Neighbor (k-NN), which is an instance-based, lazy learning algorithm. It computes distance between a test instance and the instances in the training set using a similarity or a distance measure such as Euclidean distance [4]. The Decision Tree (DT) is another type of classifier, which is constructed in a top-down, recursive and divide-and-conquer manner [1]. However, the applicability of DT is limited in text classification due to the extremely high dimensionality of the Bag-Of-Words representation of the documents. Among many different types of classifiers, Naive Bayes (NB) is one of the simplest one. It can be understood easily. It has a well-balanced training and classification time complexities; both are relatively very low. The implementation of NB algorithm is also easier than that of other classifiers. One of the most common event models of Navies Bayes classifier is called Multinomial Naïve Bayes (MNB). We used MNB and SVM to compare the new SMC classifier proposed in this study. They are by far the most commonly used classifiers for text classification.

### *B. Helmholtz principle*

According to Helmholtz principle from Gestalt theory of human perception; " an observed geometric structure is perceptually meaningful if it has a very low probability to appear in noise" [15]. This means that interesting events happens in large deviations from randomness. This can be illustrated in Figure 1. In the left hand side of Figure 1, there is a group of five aligned dots but it is not easy to notice it due to the high noise. Because of large number of randomly placed dots, the alignment probability of five dots increases. On the other hand, if we remove the number of randomly placed dots considerably, we can immediately perceive the alignment pattern in the right hand side image since it is very unlikely to happen by chance. This phenomenon means that unusual and rapid changes will not happen by chance and humans can immediately perceive them.

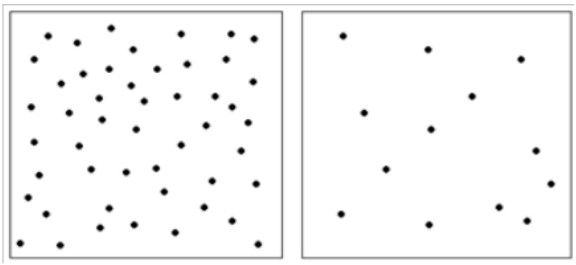

Figure 1 : Helmholtz principle [14]

## III. APPROACH

Meaning measure is based on Helmholtz principle from Gestalt theory and it is defined with following formulas. These formulas basically calculate how likely is an event in a particular context. In our case, this event is the occurrence frequency of a term in a particular context such as a document or a class of documents. In order to measure this, first Number of False Alarms (NFA) is calculated. Meaningfulness of a word occurrence frequency is proportional to NFA, which uses the frequencies of this word in the particular context and the whole dataset.

$$NFA(w,P,D) = \binom{K}{m}\frac{1}{N^{m-1}} \quad (1)$$

$$Meaning(w,P,D) = -\frac{1}{m}\log NFA(w,P,D) \quad (2)$$

$$N = \frac{|D|}{|P|} \quad (3)$$

In particular, these formulas calculate the meaning of word $w$ in partition $P$ of whole document $D$. In this study, we adopt $w$ as a word or term, $P$ represents the documents belonging to a particular class and $D$ represents the whole dataset or corpus available to us during the calculations. The word $w$ appears $m$ times in $P$ and $K$ times in $D$. $N$ is equal to the length of $D$ divided by the length of $P$ in terms of number of words

At the end of the meaning calculations, we know the meaningfulness of each word in each class. Meaning values of words in each class can be positive and negative. If the words have negative meaning value in class $P$, this means that these words are not important for the class and they are called non-meaningful words. These are ignored in the further calculations. The set of non-meaningful words can be computed with the following formula (4) [13].

$$\{w : Meaning(w,P,D) < 0\} \quad (4)$$

Calculating the meaning values of the terms in the context of each class constitutes the training phase of the SMC algorithm. Please note that this is similar to the Naïve Bayes approach where the class conditional term probabilities are calculated in the training phase. On the classification phase, when we need to assign a class label to a previously unseen document, we calculate the meaningfulness of the document for each class $P$ by summing the meaning values of the terms for that class as in (5). Next, the document is assigned to the class with the highest meaning value. This is similar to the classification phase of Naïve Bayes where the class conditional term probabilities are multiplied to calculate the class conditional document probability.

$$Meaning(d,P) = \sum_{i}^{|W|} Meaning(w,P,D) \quad \forall w \in d \quad (5)$$

## IV. EXPERIMENT SETUP

During experiments, we used WEKA machine learning toolkit [19]. We used WEKA implementations of MNB and SVM (SMO). The default parameters of the SMO are linear kernel and complexity parameter of 1. We applied 10 random trials by using a particular training set percentage and a constant 20% test set. This is quite similar to the well-known 10-fold cross validation approach. The main difference is the size of the training set. In 10-fold cross validation the size of the training set is fixed to 90%. On the other hand, in order to observe the affect of scarce training set on the performance of the proposed algorithm we start with 1% training set size and increase it up to 70% in our random trials. The performance of

classifiers is measured using accuracy metric. Accuracy is the percentage of correct classifications among the test set.

We used a Turkish dataset; 1150haber [20] and two English datasets; mini-newsgroups[1] and imdb[2]. Table I gives properties of these datasets. These datasets have a balanced class distribution. The imdb is a popular dataset used in sentiment analysis and opinion mining studies. It consist of positive and negative comments or sentiments on movies. The mini-newsgroups dataset is a subset of the 20 Newsgroups dataset, which is one of the most common datasets used in text mining studies. It consists of news-groups postings organized in 20 hierarchical categories. However, we do not use the hierarchical structure. The 1150haber on the other hand consist of newspaper articles, i.e. columns in five different categories.

We do not use any stop-words removal and feature selection methods, which lessen the effect of commonly used words and to reduce the number of features respectively. The meaning calculations can naturally omit stop words and assigns higher values to more significant and informative words [13].

TABLE I. PROPERTIES OF DATASETS

| Dataset | #Classes | #Instances | #Attributes |
|---|---|---|---|
| 1150haber | 5 | 1,150 | 6,656 |
| Mini-newsgroups | 20 | 2,000 | 12,112 |
| Imdb | 2 | 2,000 | 16,679 |

In order to observe the performance of SMC under different training set size conditions, we divide data set into two parts as training set and testing test. The selected percentage of training data ranges from 70% to 1%. This enables us to see the effect of insufficient labeled data on the performance of the new SMC classifier. In the following section, we report the average of 10 random trials in each training set percentage.

## V. EXPERIMENT RESULTS AND DISCUSSION

According to our experiments, SMC demonstrates a notable performance on datasets especially when the size of training data is small.

TABLE II. ACCURACY OF DIFFERENT CLASSIFIERS ON 1150HABER DATASET WITH VARYING TRAINING SET SIZE

| *TS%* | *SMC* | *MNB* | *SVM* |
|---|---|---|---|
| 70 | 94.26 | **94.65** | 90.65 |
| 50 | **92.96** | 92.91 | 87.61 |
| 30 | **93.39** | 93.04 | 84.65 |
| 10 | **90.70** | 85.04 | 74.96 |
| 5 | **84.70** | 70.17 | 67.22 |
| 1 | **52.35** | 38.35 | 35.43 |

Table II shows accuracies of each classifier at different training set size proportions (TS%) on 1150haber dataset. The results show that SMC outperforms other popular classifiers. Although we do not report run time measurements, we note that SMC is much faster than SVM. When training data is very low (%1), SMC is especially more effective and has much higher accuracy than other classifiers with insufficient training data.

TABLE III. ACCURACY OF DIFFERENT CLASSIFIERS ON MINI-NEWSGROUPS DATASET WITH VARYING TRAINING SET SIZE

| *TS%* | *SMC* | *MNB* | *SVM* |
|---|---|---|---|
| 70 | **84.03** | 75.93 | 68.68 |
| 50 | **82.95** | 74.53 | 65.50 |
| 30 | **81.03** | 62.00 | 60.90 |
| 10 | **69.93** | 32.33 | 40.80 |
| 5 | **64.35** | 18.85 | 30.08 |
| 1 | **48.18** | 10.95 | 19.60 |

Table III shows accuracies of each classifier at different training set size proportions on mini-newsgroups dataset. The results show that SMC outperforms other popular classifiers at all training set sizes by a considerable margin. The difference is especially visible at small training set sizes. The proposed classifier performs exceptionally well on this dataset. We speculate that this is due to the relatively larger number of classes (20 classes vs. 5 classes in 1150haber and only two classes in imdb) and the relatively smaller number of documents in each class (only 100 documents vs. 230 documents in 1150haber and 1000 documents in imdb).

TABLE IV. ACCURACY OF DIFFERENT CLASSIFIERS ON IMDB DATASET WITH VARYING TRAINING SET SIZE

| *TS%* | *SMC* | *MNB* | *SVM* |
|---|---|---|---|
| 70 | 75.08 | 81.98 | **83.28** |
| 50 | 73.98 | 80.15 | **82.20** |
| 30 | 72.70 | 78.75 | **80.30** |
| 10 | 68.63 | 74.00 | **74.80** |
| 5 | 67.78 | 69.68 | **70.80** |
| 1 | **69.65** | 59.60 | 66.45 |

Table IV shows accuracies of each classifier at different training set size proportions on imdb dataset. This dataset is used in sentiment analysis studies. It consists of positive and negative comments about movies. On this dataset, SVM has better performance results except one case. The SMC has a better performance only on the last row where we have insufficient training data (1%). We think that the relatively smaller number of classes, only two classes, makes it harder for meaning calculations to distinguish between classes. On the other hand, this dataset is a good fit for a binary classifier such as SVM. An additional observation is related to the relatively larger number of documents in each class, which gives an advantage to the traditional classifiers to extract patterns.

[1] http://archive.ics.uci.edu/ml/

[2] http://www.imdb.com/interfaces

The t-test a is statistical hypothesis test. This test is commonly used in evaluating the results of two classifiers combined with 10-fold cross validation or several fold random trial experiments like our approach. It is used in order to find out if the difference between two classifiers measured by a metric such as accuracy is statistically significant or not. Usually a threshold on P value is used for determining the significance and 0.05 is commonly used in the literature [21]. Similarly, we use this threshold and t-test to determine if the accuracy results of 10 random trials of two classifiers are statistically significant or not.

TABLE V. T-TEST OF SMC WITH DIFFERENT CLASSIFIERS ON 1150HABER DATASET WITH VARYING TRAINING SET SIZE

| *TS%* | *MNB* | *SVM* |
|---|---|---|
| 70 | - | + |
| 50 | - | + |
| 30 | - | + |
| 10 | + | + |
| 5 | + | + |
| 1 | + | + |

Table V indicates t-Test results of SCM and other classifiers for each training set sizes on 1150haber. In this table and the following tables, (+) indicates statistically significant difference (t-test $P<0.05$). The results in this table show that SMC statistically significantly outperforms SVM in all training set percentages and MNB up to 30% training set level. It is important to note that, although the t-test value is higher than the threshold for SMC and MNB, there is a visible difference between the accuracies, such as SMC reaching up to 84.03% accuracy while MNB can only reach 75.93% accuracy.

TABLE VI. T-TEST OF SMC WITH DIFFERENT CLASSIFIERS ON MINI-NEWSGROUPS DATASET WITH VARYING TRAINING SET SIZE

| *TS%* | *MNB* | *SVM* |
|---|---|---|
| 70 | + | + |
| 50 | + | + |
| 30 | + | + |
| 10 | + | + |
| 5 | + | + |
| 1 | + | + |

Table VI indicates t-Test results of SCM and other classifiers for each TS rates. SMC performs specifically well on this dataset with 20 classes. The results in this table show that SMC statistically significantly outperforms both SVM and MNB in all training set percentages. It is important to note that SMC outperforms other classifiers by a wide margin on this dataset.

TABLE VII. T-TEST OF SMC WITH DIFFERENT CLASSIFIERS ON IMDB DATASET WITH VARYING TRAINING SET SIZE

| *TS%* | *MNB* | *SVM* |
|---|---|---|
| 70 | + | + |
| 50 | + | + |
| 30 | + | + |
| 10 | + | + |
| 5 | - | + |
| 1 | + | + |

Table VII indicates t-test results of SCM and other classifiers for each TS rates. The results in this table show that SMC statistically significantly outperforms SVM and MBN in 1% training set level and although the accuracy of MNB is higher at 5% training set, it is not significant.

## VI. CONCLUSIONS AND FUTURE WORK

We propose a new text classifier named Supervised Meaning Classifier (SMC). The novel SMC classifier uses meaning measure, which is based on Helmholtz principle from Gestalt Theory. The Helmholtz Principle from Gestalt Theory suggests that if an observed geometric structure has very low probability to appear in noise, this geometric structure is perceptually meaningful. By using Helmholtz Principle from Gestalt Theory, Balinsky et al. [15] presents and defines the "meaning measure" to be used in textual, unstructured and sequential data mining applications. This measure uses the fact that interesting events appear as large deviations from randomness. Previously, we adopted and applied meaning measure for supervised and unsupervised feature selection [8][9][10] and as a semantic kernel [25]. The novel classifier proposed in this study uses the supervised variant of the meaning measure [8] [9], which calculates the meaning of terms in the context of classes of documents. In SMC, meaningfulness of terms in the context of classes are calculated and used for classification of a document. Our detailed experimental results show that the SMC outperforms other popular text classification algorithms of MNB and SVM especially when the training data is insufficient and the number of classes is large. Also our observations show that SMC much faster than SVM and its speed is comparable with MNB.

In the future, we would like to further analyze the theory behind the class based meaning calculations and improve SMC by incorporating more information from the semantic relations between terms. We also would like to shed light into several issues such as the role of terms with negative meaning values and the terms that exist only in test set.

## ACKNOWLEDGMENT

This work is supported in part by The Scientific and Technological Research Council of Turkey (TÜBİTAK) grant number 111E239. Points of view in this document are those of

the authors and do not necessarily represent the official position or policies of the TÜBİTAK.

## REFERENCES


[1] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann, Third Edition,2012

[2] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features (pp. 137-142). Springer Berlin Heidelberg.

[3] Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In Machine learning: ECML-98 (pp. 4-15). Springer Berlin Heidelberg.

[4] Weiss, S., Kasif, S. and Brill, E. (1996) Text classification in USENET newsgroup: a progress report. In AAAI Spring Symp. on Machine Learning in Information Access Technical Papers, Palo Alto, March 1996.

[5] Lewis, D. D. and Ringutte, M. (1994) A comparison of two learning algorithms for text categorization. In Third Annual Symp. on Document Analysis and Information Retrieval, Las Vegas, NV, pp. 81–93.

[6] Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003, August). Tackling the poor assumptions of naive bayes text classifiers. In ICML (Vol. 3, pp. 616-623).

[7] Li, Y. H., & Jain, A. K. (1998). Classification of text documents. The Computer Journal, 41(8), 537-546.

[8] Tutkan, M., Ganiz, M.C., Akyokuş, S.(submited : 2015). Helmholtz Principle based Supervised and Unsupervised Feature Selection Methods for Text Mining, Information Processing & Management Journal.

[9] Tutkan, M., Ganiz, M.C., Akyokuş, S.(2014) Metin Sınıflandırma için Yeni Bir Eğitimli Anlamsal Özellik Seçimi Yöntemi. ASYU 2014 (Akıllı Sistemlerde Yenilikler ve Uygulamalar Sempozyumu) , Ekim 9-10, İzmir Katip Çelebi Üniversitesi, İzmir, Türkiye.

[10] Tutkan, M., Ganiz, M.C., Akyokuş, S.(2014). Metin Sınıflandırma için Yeni Bir Eğitimsiz Anlamsal Özellik Seçimi Yöntemi. ELECO 2014 (Elektrik - Elektronik, Bilgisayar ve Biyomedikal Mühendisliği Sempozyumu), 27-29 Kasım 2014, Bursa, Türkiye.

[11] Desolneux, A., Moisan, L., & Morel, J. M. (2007). From gestalt theory to image analysis: a probabilistic approach (Vol. 34). Springer

[12] Dadachev, B., Balinsky, A., & Balinsky, H. (2014, September). On automatic text segmentation. In Proceedings of the 2014 ACM symposium on Document engineering (pp. 73-80). ACM.

[13] Dadachev, B., Balinsky, A., Balinsky, H., & Simske, S. (2012, September). On the helmholtz principle for data mining. In Emerging Security Technologies (EST), 2012 Third International Conference on (pp. 99-102). IEEE.

[14] Balinsky, A. A., Balinsky, H. Y., & Simske, S. J. (2010, September). On Helmholtz's principle for documents processing. In Proceedings of the 10th ACM symposium on Document engineering (pp. 283-286). ACM.

[15] Balinsky, A., Balinsky, H., & Simske, S. (2011b). On the Helmholtz principle for data mining. Hewlett-Packard Development Company, LP.

[16] Balinsky, A., Balinsky, H., & Simske, S. (2011c, October). Rapid change detection and text mining. In The 2nd IMA Conf. on Mathematics in Defence, Defence Academy, UK.

[17] Balinsky, H., Balinsky, A., & Simske, S. (2011a, October). Document sentences as a small world. In Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on (pp. 2583-2588). IEEE.

[18] Balinsky, H., Balinsky, A., & Simske, S. J. (2011d, September). Automatic text summarization and small-world networks. In Proceedings of the 11th ACM symposium on Document engineering (pp. 175-184). ACM.

[19] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.

[20] Amasyalı, M. F., & Beken, A. (2009). Türkçe kelimelerin anlamsal benzerliklerinin ölçülmesi ve metin sınıflandırmada kullanılması. In IEEE Signal Processing and Communications Applications Conference, SİU-2009.

[21] O'Mahony, M. (1986). Sensory evaluation of food: statistical methods and procedures (Vol. 16). CRC Press.

[22] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Steinberg, D. (2008). Top 10 algorithms in data mining. Knowledge and Information Systems, 14(1), 1-37.

[23] Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. ACM Comput. Surv. 34, 1 (March 2002), 1-47.

[24] Charu C. Aggarwal, ChengXiang Zha. (2012) A Survey of Text Classification Algorithms. Mining Text Data, pp. 163-222, Springer.

[25] Altınel, B., Ganiz, M.C., Diri, B., (2015). A Corpus-Based Semantic Kernel for Text Classification by Using Meaning Values of Terms. Engineering Applications of Artificial Intelligence , Vol. 43, pp. 54-66