# Improving Twitter Sentiment Classification Using Term Usage and User Based Attributes

Selim Akyokuş[1], Murat Can Ganiz[2], Cem Gümüş[3]
Dogus University[1], Marmara University[2], Dogus University[3]
Istanbul, Turkey
e-mail: sakyokus@dogus.edu.tr[1], murat.ganiz@marmara.edu.tr[2], 201091001@dogus.edu.tr[3]

***Abstract*—With the rapid growth of the Internet and the increase in the use of mobile devices, social media has grown rapidly in recent years. Without using appropriate representation techniques, processing methods and algorithms, it is difficult to get patterns, trends and opinions that are of interest to companies, organizations and individuals. Sentiment classification, which is one of the most popular mining tasks on the textual part of the social media data, aims to classify comment texts by their polarity. Textual features such as terms, n-grams combined with the NLP techniques are commonly used for this task. Our aim in this study is to see the effect of additional features on Twitter sentiment classification that are extracted from structured data related to the tweets and the Twitter users associated with these tweets. In addition to the use of terms in tweets as features i.e. traditional bag-of-words model, we employed tweet term usage based attributes along with Twitter user based attributes and showed that these additional attributes increase the accuracy of class substantially.**

*Keywords-component; Sentiment Analysis; Sentiment Classification; Machine Learning; Feature Engineering; Feature Extraction;*

## I. INTRODUCTION

People's desire to share ideas, opinions and suggestions using social media has enabled the collection of huge amounts of data on the Internet. The raw data kept in social media environments must be preprocessed, represented, and analyzed in order to extract important patterns and trends. Typos, heavy use of slang, abbreviations, emotional expressions and the use of informal - daily conversation language make it difficult to work on the textual part of the social media data.

Twitter is one of the most widely used social media environments that have attracted many researchers for sentiment analysis. Sentiment analysis on Twitter data is more difficult than traditional textual documents due to characteristics of Twitter data. Twitter allows users to post messages of at most 140 characters. Because of this limitation, users tend to abbreviate words, use special characters and acronyms. The majority of messages are about current news and events in a conversional style. Although Twitter messages are short, the number of messages and different terms used in messages about a topic can be very high. This causes high dimensionality and sparsity on Twitter data sets.

The Twitter system allows researchers to collect tweets by using publicly available Application Programming Interfaces (APIs). Using the API, tweets about specified keywords and phrases can be obtained as a stream. Many studies have been done on Twitter messages by collecting data with this API. Examples of these studies include studies predicting outbreaks [1], examining medications and their unknown side effects [2], estimating changes in human perception over time [3], and perceptual analysis on the tweets of tourists coming to a tourist destination [4]. In the field of emotion analysis, although there are many studies for Twitter data written in English [5]-[8], a limited number of studies have been done for Turkish [9][10].

In this study, firstly data was collected from Twitter with a custom crawler application. The Web application was developed for data labelling. Tweets were shown to Dogus University students by this application. Of these tweets, all content is only in Turkish were labeled by the Dogus University students. After this, we preprocessed the labeled Twitter data. The preprocessing step included removal of stopwords, normalization of some terms, tokenization, and formation of term-document (tweet) matrix with Term Frequency-Inverse Document Frequency (TF-IDF) [18] weighting. We also computed several term and user statistics as additional features to be added to the term-document matrix. The additional features included user tweet counts and tweet term usage rate information. Balanced and unbalanced data sets were prepared with these collected data. Several classification algorithms from machine learning domain have been applied on to these datasets and the effects of the additional features have been investigated.

This paper is organized as follows: Section II presents general aspects of data preparation. In Section III, we show the results of experiments. Last Section summarizes our contribution.

## II. DATA SET

### *A.* Data collection and storage

To collect Twitter data, a Java application has been developed using Twitter API. This application obtained tweets written in all languages from the Twitter system. We collected tweets written in all languages. The collected

 ISBN: 978-1-61208-566-1 

Twitter records were saved in the tables created in the PostgreSQL [13] relational database.

*B.* Data labelling

Data labelling manually is a tedious work and requires many people. That's why we chose students from our university. A Web application has been developed using the ZK framework [14], Spring framework [15], Hibernate [16] and Java [17] to label the collected data so that it can be used in classification. Tweets about Turkish companies operating in banking, telecom companies, universities and mobile phone device brands are shown to Dogus University students by the Web application. Of these tweets, mixed type of tweets were not labelled (e.g., half Turkish, half English). Only those all content is with Turkish were labeled by the students as positive, negative and neutral by using this application. Our study, each tweet labelled by a single student. Depending on the content of each tweet was labelled by students according their opinion and feelings. Within the scope of this study, 20204 tweets were labelled. Table I shows the number of labelled tweets.

- TT-BC: Tweets about banking.
- TT-TC: Tweets about telecom companies.
- TT-US: Tweets about universities.
- TT-PB: Tweets about mobile phone device brands.

TABLE I. LABELLED TWEET DETAILS

| Tweet Topic | Type | | | |
|---|---|---|---|---|
| | Positive | Negative | Neutral | Total |
| TT-BC | 1451 | 4603 | 1997 | 8051 |
| TT-TC | 2226 | 2738 | 884 | 5848 |
| TT-US | 1429 | 2230 | 1332 | 4991 |
| TT-PB | 586 | 322 | 406 | 1314 |
| **Total** | **5692** | **9893** | **4619** | **20204** |

*C.* Data preprocessing

There are some irrelevant terms and character sequences in Twitter messages that are not valuable or informative for classification tasks. Messages posted by Twitter users may include the following irrelevant terms.

- User names starting with "@" character ,
- Hashtags starting with "#" character ,
- Emotion expressions and
- URLs

Some data cleansing and preprocessing work were performed to remove these terms so that more effective results can be obtained in experiments. In addition, repeated messages shared by a person, messages containing only a URL, hashtag, special character, number and emotion expressions were deleted before the preprocessing steps. During the preprocessing step, Twitter messages about telecom companies were processed as described below.

*1)* Tokenize strings: It is a process that tries to tokenize messages and get meaningful data from them. The following operations have been applied:

- The URL, hashtag, usernames and special characters in the messages have been deleted.
- The contents of the messages have been converted to lowercase and all characters outside the letters have been deleted.

*2)* Stemming (Root finding): Stemming is a means for grouping words with a similar meaning together. In stemming, stemming algorithms transform inflected words to their word stem, or root form. For this purpose, the Zemberek library [12] was used to find the roots of Turkish words.

*3)* Correction of erroneous terms: It is a process that aims to correct terms that were mistakenly written in messages. The propositional function of the Zemberek library [12] was used for this process.

*4)* Deletion of repeated terms: It is a process that aims to reduce the size of characters and the correction of repetitive letters in the terms used in messages. In this study, repetitive letters in terms were deleted (e.g., Haapppyyyyy).

*5)* TF-IDF [18] weighting: In TF-IDF weighting scheme, a weight of each term in document is computed. Each weight represents the importance of a term inside a document [10]. TF-IDF was calculated for each term as follows:

$$TF(t,d) = 1 + \log_{10} f_d(t) \quad (1)$$

$$IDF(t,D) = \log_{10}\left(\frac{|D|}{df(t)}\right) \quad (2)$$

$$TF\text{-}IDF(t,d,D) = TF(t,d).\ IDF(t,D) \quad (3)$$

Where,

$f_d(t)$ : Frequency of term t in document (tweet)
d : Document in corpus
$df(t)$ : The number of tweets that contain term $t$
$D$ : Corpus of documents (tweets)
$|D|$ : Total number of tweets in corpus

*6)* Calculation of tweet term usage statistics: Positive, negative, neutral and total tweet term usage rates were calculated for use in experiments. Equations about term statistics are our equations. The values calculated for each of these tweets were added as attributes to term-document matrix.

- The Tweet Term Usage Rate is calculated as follows (4):

 ISBN: 978-1-61208-566-1 

$$WT(d_i) = \sum_{j=0}^{|d_i|} \log_{10} \frac{|DT|}{f_D(d_{i,j})} \quad (4)$$

Where,

$|DT|$ : Total number of terms in corpus
$d_i$ : Document (tweet) $i$ in tweet corpus
$|d_i|$ : The number of terms in document $i$
$d_{i,j}$ : $j^{th}$ term in document $i$
$f_D(t)$ : Frequency of term $t$ in all tweets in corpus

- The Tweet Term Positive Usage Rate is calculated as follows (5):

$$WP(d_i) = \sum_{j=0}^{|d_i|} \log_{10} \frac{|DP|}{f_{DP}(d_{i,j})} \quad (5)$$

Where,

$|DP|$ : Total number of terms in positive tweets corpus
$f_{DP}(t)$ : Frequency of term $t$ in positive tweets in corpus

- The Tweet Term Negative Usage Rate is calculated as follows (6):

$$WN(d_i) = \sum_{j=0}^{|d_i|} \log_{10} \frac{|DN|}{f_{DN}(d_{i,j})} \quad (6)$$

Where,

|DN| : Total number of terms in negative tweets corpus
$f_{DN}(t)$ : Frequency of term $t$ in negative tweets in corpus

- The Tweet Term Neutral Usage Rate is calculated as follows (7):

$$WR(d_i) = \sum_{j=0}^{|d_i|} \log_{10} \frac{|DR|}{f_{DR}(d_{i,j})} \quad (7)$$

Where,

$|DR|$ : Total number of terms in neutral tweets corpus
$f_{DR}(t)$ : Frequency of term $t$ in neutral tweets in corpus

*7)* Finding user statistics (tweet counts): Positive, negative, neutral and total tweet counts of users were calculated for use in experiments. The values calculated for each user have been added as attributes to term-document matrix.

$U_t(i)$ : Total number of tweets posted by user $i$
$U_p(i)$ : Total number of positive tweets posted by user $i$
$U_n(i)$ : Total number of negative tweets posted user $i$
$U_r(i)$ : Total number of neutral tweets posted by user $i$

*D.* Data Set Preparation

After data cleansing and preprocessing on tweets, several datasets were prepared by taking tweets about telecom companies for use in experiments. A Java [17] application has been developed to prepare data sets. Using this application, two types of data sets were created for the experiments: balanced and unbalanced. The classification on the balanced data set is more successful than expected. For this reason, we wanted to see the differences by preparing a balanced and unbalanced dataset. In balanced data sets, the number of instances in each class is the same. The number of instances in unbalanced data sets and the balanced data sets are shown in the Table II. In addition, we used four different representation methods:

- TF-IDF: Term-document matrix includes entries where each value is weighed using TF-IDF method.
- TF-IDF + US: User statistics features added to TF-IDF matrix
- TF-IDF + TS: Term statistics features are added to TF-IDF matrix
- TF-IDF + TS + US: Both term statistics and user statistics features are added to TF-IDF matrix.

TABLE II. DATA SET DETAILS

| Data Set Type | Type | | | |
|---|---|---|---|---|
| | Positive | Negative | Neutral | Total |
| Unbalaced | 1272 | 1140 | 504 | 2916 |
| Balanced | 504 | 504 | 504 | 1512 |

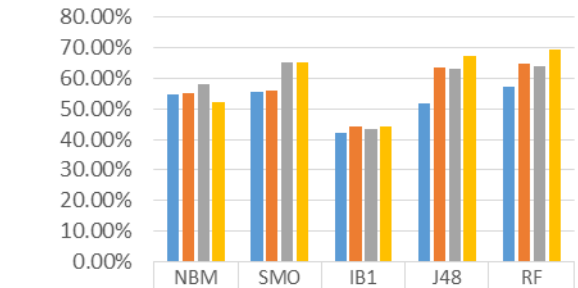
## III. EXPERIMENTS

We used Weka [11] for sentiment classification with default Weka [11] parameters. Weka [11] is a widely used tool written in Java [17] for data mining research. It includes a collection of machine learning algorithms for data mining tasks. Naive Bayes Multinomial (NBM), Random Forest (RF), Sequential Minimal Optimization (SMO), Decision Tree (J48) and 1-Nearest Neighbors (IB1) algorithms [19]-[21] are used for sentiment classification in our experiments. 10-fold cross-validation and repeated holdout methods were

used as accuracy estimation methods. In repeated holdout method, the data set was randomly separated 10 times into two sets: 80% for training and 20% for testing. Then, average accuracies of classifiers were computed using 10 tests.

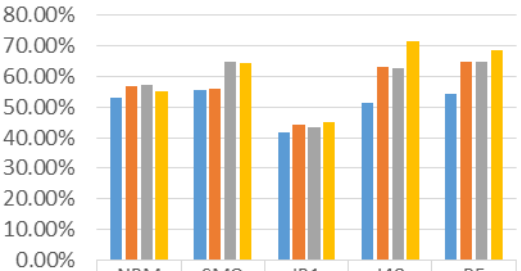
The experiment with 10-fold cross validation on unbalanced data set is shown in Figure 1. The first row shows accuracies of different classification algorithms using only TF-IDF weighting method. In the second row, we added four features $U_t$, $U_p$, $U_n$ and $U_r$ involving user statistics to see their effects. The third row shows the effect of term statistics obtained by formulas (4)-(7). The last row (labeled TF-IDF + TS + US) displays the accuracies of algorithms the data set which includes TF-IDF weighting, term statistics and user statistics. Figure 2 shows accuracies of classifiers using repeated holdout method with 10 repetitions. The best performance results are obtained with decision tree (J48) and random forest algorithms. From the last two columns, we can observe that both user statistics and term statistics features increase the performances of classifiers. The best accuracy 71.70% is obtained by applying J48 algorithm on the data set that includes all TF-IDF weighting, term statistics and user statistics.

The experiment with 10-fold cross validation on balanced data set is shown in Figure 3. Figure 4 shows the results of experiments using repeated holdout method with 10 repetitions on balanced datasets. As it can be seen in Figures 3 and 4, balanced data sets produce better performance results than unbalanced datasets. Again, the better performances are obtained by applying J48 and RF algorithms. The best accuracy 80.22% is achieved with J48 algorithm using all features TF-IDF + TS + US. Classifications accuracies in references [22]–[24] are 76%, 45% and 64% respectively. Our accuracy results are 71.70 and 80.22%. Although it is difficult to compare results of different research studies that use different data sets, we obtained relatively better results than the most of other research studies.



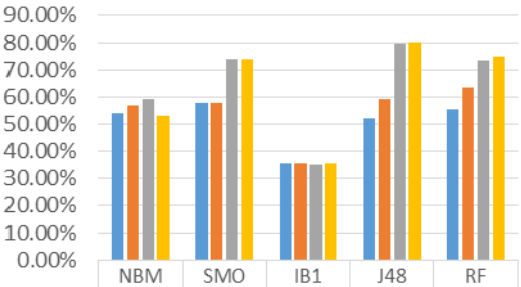

| | NBM | SMO | IB1 | J48 | RF |
|---|---|---|---|---|---|
| TF-IDF | 54.90% | 55.76% | 42.25% | 51.61% | 57.20% |
| TF-IDF+US | 54.97% | 55.86% | 44.38% | 63.55% | 64.75% |
| TF-IDF+TS | 57.99% | 65.12% | 43.31% | 63.31% | 64.09% |
| TF-IDF+US+TS | 52.23% | 65.33% | 44.38% | 67.15% | 69.38% |

Figure 1. Unbalanced data set experiments with 10-fold cross-validation



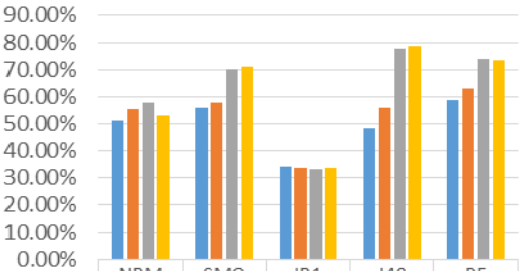

| | NBM | SMO | IB1 | J48 | RF |
|---|---|---|---|---|---|
| TF-IDF | 53.00% | 55.75% | 41.68% | 51.46% | 54.20% |
| TF-IDF+US | 56.95% | 56.09% | 44.25% | 63.29% | 64.67% |
| TF-IDF+TS | 57.12% | 64.84% | 43.40% | 62.61% | 65.01% |
| TF-IDF+US+TS | 55.23% | 64.32% | 45.11% | 71.70% | 68.61% |

Figure 2. Unbalanced data set experiments with repeated holdout method




| | NBM | SMO | IB1 | J48 | RF |
|---|---|---|---|---|---|
| TF-IDF | 54.17% | 57.67% | 35.59% | 52.18% | 55.62% |
| TF-IDF+US | 56.88% | 57.67% | 35.59% | 59.19% | 63.56% |
| TF-IDF+TS | 59.19% | 73.74% | 35.25% | 79.63% | 73.41% |
| TF-IDF+US+TS | 53.04% | 73.88% | 35.45% | 80.22% | 74.87% |

Figure 3. Balanced data set experiments with 10-fold cross-validation




| | NBM | SMO | IB1 | J48 | RF |
|---|---|---|---|---|---|
| TF-IDF | 51.36% | 56.09% | 34.24% | 48.42% | 58.61% |
| TF-IDF+US | 55.65% | 57.98% | 33.85% | 56.03% | 62.84% |
| TF-IDF+TS | 57.78% | 70.04% | 33.27% | 77.82% | 73.93% |
| TF-IDF+US+TS | 53.11% | 70.82% | 33.46% | 78.60% | 73.35% |

Figure 4. Balanced data set experiments with repeated holdout

## IV. CONCLUSION AND FUTURE WORK

In this study we extract additional features for Twitter sentiment classification form tweets and user information. In addition terms in a bag-of-words model weighted with TF-IDF, we also derived 8 new features about user and term usage statistics. To observe the effect of these additional features on Twitter sentiment classification we collect and label tweets and after that conduct several experiments with different conditions using several different machine learning algorithms.

Experiments show that the additional features considerably increase the performance of classifiers, especially when the dataset has a skewed class distribution. As future work, we plan to apply semi-supervised algorithms used in situations where most of the samples are unlabeled and there exists a small number of labeled samples.

## REFERENCES


[1] Martin Szomszor, Patty Kostkova, and Ed De Quincey. "# swineflu: Twitter predicts swine flu outbreak in 2009." 3rd International ICST Conference on Electronic Healthcare for the 21st Century (eHEALTH2010). pp. 18-26, 2012.

[2] Jiang Bian, Umit Topaloglu, and Fan Yu. "Towards large-scale twitter mining for drug-related adverse events." Proceedings of the 2012 international workshop on Smart health and wellbeing. ACM, pp. 25-32, 2012.

[3] Le Thanh Nguyen, Pang Wu, William Chan, Wei Peng and Ying Zhang, "Predicting collective sentiment dynamics from time-series social media." Proceedings of the first international workshop on issues of sentiment discovery and opinion mining. ACM, p.6, 2012.

[4] William B. Claster, Hung Dinh, and Malcolm Cooper. "Naïve Bayes and unsupervised artificial neural nets for Cancun tourism social media data analysis." Nature and Biologically Inspired Computing (NaBIC), 2010 Second World Congress on. IEEE, pp. 158-163, 2010.

[5] Alec Go, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford 1.12 (2009). Mishne G., Natalie S. "Predicting Movie Sales fromBlogger Sentiment." AAAI Spring Symposium: Computational Approachesto Analyzing Weblogs. 2006.

[6] Alexander Pak, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." LREc. Vol. 10. No. 2010. 2010.

[7] Dmitry Davidov, Oren Tsur, and Ari Rappoport. "Enhanced sentiment learning using twitter hashtags and smileys." Proceedings of the 23rd international conference on computational linguistics: posters. Association for Computational Linguistics, pp. 241-249, 2010.

[8] Mesut Kaya, Guven Fidan, and Ismail Hakki Toroslu. "Sentiment analysis of turkish political news." Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01. IEEE Computer Society, pp. 174-180, 2012.

[9] Cumali Türkmenoglu, and Ahmet Cüneyd Tantug. "Sentiment analysis in Turkish media." Proceedings of Workshop on Issues of Sentiment Discovery and Opinion Mining, International Conference on Machine Learning (ICML), Beijing, China. 2014.

[10] Gerard Salton, and Christopher Buckley. "Term-weighting approaches in automatic text retrieval." Information processing & management 24.5 (1988): 513-523.

[11] Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal. "Data Mining: Practical machine learning tools and techniques." Morgan Kaufmann Publishing, 2016.

[12] Ahmet Afsin Akın, and Mehmet Dündar Akın. "Zemberek, an open source NLP framework for Turkic languages." Structure 10 (2007): 1-5. 2007.

[13] PostgreSQL Relational Database, www.postgresql.org

[14] ZK Enterprise Java Web Framework, www.zkoss.org

[15] Spring Framework, www.spring.io

[16] Hibernate ORM(Object Relational Mapping), www.hibernate.org

[17] Java Programming Language, www.java.com

[18] Mingyoug Liu, and Jiangang Yang. "An improvement of TFIDF weighting in text categorization." International Proceedings of Computer Science and Information Technology (2012): 44-47.

[19] Jiawei Han, and Micheline Kamber. "Data Mining: Concepts and Techniques." Morgan Kaufmann Publishing, 2006.

[20] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. "Introduction to Data Mining." Addison-Wesley Publishing, 2006.

[21] Mehmed Kantardzic. "Data Mining: Concepts, Models, Methods, and Algorithms." John Wiley & Sons Publishing, 2003.

[22] A. Gural Vural, B. Barla Cambazoglu, Pinar Senkul, and Z. Ozge Tokgoz. "A framework for sentiment analysis in Turkish: Application to polarity detection of movie reviews in Turkish." Computer and Information Sciences III. Springer London, 2013. 437-445.

[23] Mehmet Ulvi Şimşek, and Suat Özdemir. "Analysis of the relation between Turkish twitter messages and stock market index." Application of Information and Communication Technologies (AICT), 2012 6th International Conference on. IEEE, 2012.

[24] M. Fatih Amasyalı. "Active learning for Turkish sentiment analysis." Innovations in Intelligent Systems and Applications (INISTA), 2013.


 ISBN: 978-1-61208-566-1 