Journal of Modern Technology and Engineering
Vol.4, No.2, 2019, pp.109-121

Jomard Publishing

SENTIMENT ANALYSIS ON TWITTER DATA FOR AZERBAIJANI LANGUAGE

Huseyn Hasanli[1], **Burak Ordin**[1*], **Samir Rustamov**[2,3]

[1]Department of Mathematics, Faculty of Science, Ege University, Izmir, Turkey
[2]School of Information Technology and Engineering, Ada University, Baku, Azerbaijan
[3]Institute of Control Systems, Baku, Azerbaijan

Abstract. The paper is devoted to the investigation of sentiment analysis of Twitter texts in Azerbaijani language. Data cleaning and annotation techniques for Azerbaijani twits are described in the paper. Different machine learning approaches including Naive Bayes, Support Vector Machines and Maximum Entropy applied for the polarity classification of Azerbaijani twits. The achieved results from machine learning techniques and Rule based algorithms have been compared and analyzed in the paper. The suggested approaches can be applied to other languages in Turkish languages group.

Keywords: Sentiment analysis, Twitter, machine learning, lexicon based approach, classification, Azerbaijani language.

Corresponding author: Burak Ordin, Department of Mathematics, Faculty of Science, Ege University, Izmir, Turkey, e-mail: *burak.ordin@ege.edu.tr*

Received: 01 June 2019; Accepted: 03 July 2019; Published: 07 August 2019.

1 Introduction

In real life, people have a different opinion about different issues. The ideas of people can be shaped after interaction between people, interaction through social media, environmental influences and other situations. Although there are questionnaires and other methods to find out the opinions of people on any subject, these methods are not very practical due to the rapidly increasing data-information around each individual. With the spread of internet and social sharing platforms, a virtual life has been created on these platforms and people express their ideas, discuss a wide range of different topics, have new ideas and have other insights. Today, Twitter platform is widely used in social media to express people's ideas. Many types of research and projects are conducted on data collected from Twitter using natural language processing and data mining methods. As a result, users share their opinions and emotions on Twitter and generate large amounts of data. These tweets are reviewed by companies to help customers improve their services and produce quality products.

Sentiment analysis is a sub-discipline within computational linguistics and data mining. The main purpose of sentiment analysis is to discover people's mood, behavior and ideas from textual data. With the increase of microblogging services, public data of these services; data targeting different sociological fields such as politics, economics, and finance are used in sentiment analysis studies (Eliaçık & Erdogan, 2015). Sentiment Analysis is an interdisciplinary field that borrows techniques from natural language processing, text analytics, and computational linguistics to extract subjective information (Duwairi et al., 2014).

Though opinions are easy for people to understand, it is not so easy for a computer to understand at the same level of understanding. The said concept of opinion consists of the following components: Target entity, Feature and appearance of the target entity, Sentiments

(positive, negative or neutral), Opinion holder, Time (Appel et al., 2015).

Name_1: @screan_name_1 "Butun izleyenleri Hemreylik Gunu ve Yeni il munasibetile tebrik edirem. Arzu edirem 2018-ci il her birinize sevinc, bol ruzi ve ugur getirsin! (I congratulate all followers on the occasion of Solidarity Day and New Year. I wish you happiness, plenty of food and success in 2018!)" (23:10 - 29 12-2017).

If we look at the example tweet given above: target entity "Hemreylik Gunu ve Yeni il (Solidarity Day and New Year)", feature or appearance "sevinc (happiness)", "bol ruzi (plenty of food)", "ugur (success)", opinion holder "Name_1", opinion expressed time "(23:10 - 29-12-2017)" and we can say that tweet's sense of value is positive because all the emotion words are positive.

In our study, we examine how microblogs can be used for sentiment analysis. We display that how to use Twitter as a corpus for sentiment analysis and opinion mining. Different classification algorithms are applied on Twitter data and results are compared.

The rest of the paper is organized as follows: Chapter 2, provides research literature on the proposed system. Chapter 3 discuss the various sentiment analysis techniques and the corresponding algorithms. Chapter 4 describes the collection of data from Twitter and preprocessing this data. Chapter 5 has present system developed approach. Chapter VI presents the results and discussion. Finally, Chapter 6 draws the conclusions of the work.

2 Literature review

This section briefly investigates previous studies on sentiment analysis, which is a sub-area of text mining.

Pang et al. (2002) examined the sentiment classification of film reviews as a specific example of a text categorization problem on a subject basis. To do this, three classification algorithms: Naive Bayes, Maximum Entropy and Support Vector Machines were used on the problem and the results were compared.

Sentiment analysis of text for Azerbaijani language had been investigated by Aida-zade et al. (2013). Multi machine learning algorithms had been applied for news classification in Azerbaijani language in Suleymanov & Rustamov, (2018); Suleymanov et al. (2018); Aida-zade et al. (2018).

The problem of extracting feature-based summaries of customer reviews of products sold online is studied by Hu & Liu (2004). Some properties of the problem generally refer to product features and functions. In the article, the problem is identified as three sub-tasks: (1) describe the characteristics of the product in which customers express their views; (2) identify review sentences that give positive or negative views for each feature; and (3) produce a summary using the discovered information.

The article published by Shailesh Kumar Yadav focused on sensitivity classification, classification techniques and which tools are available for sensitivity analysis. In this area, it is stated that there are difficulties in discovering the polarity of the complex sentence, extracting the idea sentences and their features in different sentences and extracting many views from the same document (Yadav, 2010).

Lin et al. (2010) presented a text classification model of self-initiated learning to design a cluster-based text classification algorithm to reduce the size of the training set and increase the efficiency of the classification application.

Smith (2011) aimed to develop machine learning techniques to recognize emotions that are interwoven into the text of a document.

Kumar et al. (2012) resorted to a hybrid approach using both corpus and dictionary-based methods to determine the semantic orientation of opinion words in tweets. A case study is presented to demonstrate the use and effectiveness of the proposed system.

Mahmood et al. (2013) analyzed the effect of published tweets in predicting the winner of the 2013 election in Pakistan. They used the Rapid Miner tool (http://rapidi.com/) to examine three different standard prediction models: the CHAID decision tree, Naive Bayes, and Support Vector Machine.

Saravan and his colleagues developed a method for associating tweets sent by users to their locations (Saravanan et al., 2013).

Rustamov et al. (2013) applied Adaptive Neuro Fuzzy Inference System (2013a), Hidden Markov Models (2013b) and Hybrid models (2018) for detection of subjectivity analysis. Same techniques had been applied for document level sentiment analysis.

The article published by Shital and Jeevan uses Twitter data to learn customers' approaches and opinions. This article illustrates the approach that helps service provider companies or industries find the views on their products, services or offers. For this purpose, tweets were taken first and then tweeted as positive, negative and neutral using Stanford NLP (Shital & Jeevan, 2017).

Supervised classification techniques based on mathematical optimization were used by Sati & Ordin (2018). They chose the algorithm of polyhedral conic functions as a supervised classification function. "Mood Blogger", a real-world data set, was used to test the proposed algorithms.

In the article published by Rane and Kumar, multi-class sentiment analysis was performed on a data set consisting of 6 major US Airlines tweets. This study began with preprocessing techniques used to clean up tweets and then showed them as vectors using the concept of deep learning (Doc2vec) to perform an analysis at the phrase level. The analysis was performed using 7 different classification strategies: Decision Tree, Random Forest, SVM, K-Nearest Neighbors, Logistic Regression, Gauss Naive Bayes and AdaBoost (Rane & Kumar, 2018).

Velioglu et al. (2018) investigated the problem of sentiment analysis in emoji for positive, negative and neutral tweets in the Turkish language. In this paper, two different methods based on word bag and FastText were used. FastText is an open-source library developed by Facebook AI Research (FAIR) to learn word and sentence classifications. In this study, firstly the word bag approach was applied as a simple and effective method. Then, FastText method was applied and it was shown that there was no significant difference between the two models.

In another paper, Chory et al. (2018) discussed the application of sentiment analysis to analyzing public satisfaction with the services of a telecommunications operator in Indonesia. Pre-classification accuracy was increased by using TF-IDF weighting, POS Labeling, and Negative management together with SVM, one of the machine learning methods, in order to make the sentiment classification system relate to the level of user satisfaction with the data service.

3 Sentiment analysis methods

Sentiment classification techniques can be examined in three parts as lexicon based approach, machine learning based approach and hybrid approach. Machine learning approach implements famous ML algorithms and uses language features. As the lexicon based approach implies, the main sources of the sentiment analysis process such as emotion word dictionary or emotion sentence dictionaries are used. The hybrid approach combines both approaches and is very important in this approach with a glossary of sentiment that plays a key role in most of the methods (Medhat et al., 2014).

The machine learning approach is more practical than other approaches in opinion mining because it has fully automated applications and can manage large web data collections. Machine Learning techniques use a training set and a test set for classification. The training set contains input features vectors and corresponding class labels. Using this training set, a classification model is developed that tries to classify input features vectors to the corresponding class labels. Next, a test set is used to validate the model by estimating the class labels of features vectors that are not seen.

The methods of sentiment classification based on machine learning can be divided into three types: supervised, unsupervised and semi-supervised learning methods (Madhoushi et al., 2015).

Constructs a model based on labeled experimental data as input in supervised classification techniques (Smith, 2011).

Supervised learning is divided into two main categories: classification and regression.

Classification is the process of finding or exploring a model (function) that helps to divide data into multiple categorical classes. Output variables are often called labels or categories. In the classification technique, the algorithm learns from the data input given to it and then uses this learning to classify new observations. This data set can be two classes or multiple classes. Classification issue requires items to be categorized according to historical data. There are many different popular supervised classification methods. The following machine learning algorithms have been widely used and provide average accuracy across most domains and different data types (Desai & Mehta, 2016). Some of the commonly used supervised learning methods in the literature are briefly described below.

3.1 Naive Bayes

The Naive Bayes classifier is the simplest and most commonly used classifier. This classifier is also a simple model that works well in text categorization. Bayes theorem is used to estimate the likelihood of a particular input belonging to a particular class. When a tweet is given, it is calculated as the probability of each class.

$$P(label\ t|) = \frac{P(t|label) \times p(label)}{p(t)},$$

where the label is a specific class and t is tweet we want to classify.

t is input to the machine learning classification. This input includes all tweet words or emotions words.

$P(label)$: the prior probability of a label, or the probability of a random property identifying the label.

$P(t)$: represents prior probability of feature i.e the probability that feature will have certain values

$P(t|label)$: represents posterior probability of feature conditioned on label i.e.

To simplify the classification work, the Naive Bayes classifier also makes a "naive" assumption, that is, all features are independent of each other.

Thus, the above equation can be rewritten as follows:

$$P(label\ t|) = \frac{P(label) \times P(f1|label) \times P(f2|label) ... \times P(fn|label)}{P(t)}.$$

Each of $f1, f2, \ldots fn$ shows a feature. Although the naive assumption is never real, the Naive Bayes classification results can be quite good.

In this case, it is attempted to predict the likelihood of a tweet's content being positive or negative. There is a column with positive and negative opinions in the training data prepared for this purpose. Example of applying the Bayesian Rule to classify the "imtahanlardan nece qorxuram (I'm afraid of the exams)" tweet written about "Exam". The classification is called "naive" because we will assume that each word of the written tweet is independent.

$$P(n|imtahanlardan\ nece\ qorxuram) =$$

$$= \frac{P(imtahanlardan\ nece\ qorxuram\ n|) \times P(nece|n) \times P(qorxuram|n) \times P(n)}{P(imtahanlardan) \times P(nece) \times P(qorxuram)}.$$

Applying Bayesian rules has enabled us to significantly simplify our solution. To solve the above equation, the probability of each event will be calculated.

$P(''imtahanlardan''|n)$: The total number of "imtahanlardan" appear in negative tweets can be defined by dividing the total number of negative tweets.

$P(''n'') = P(''negative'')$: Divided by the total number of words in negative tweets and the total number of words in the training data.

$P(''imtahanlardan'')$: The total number of "imtahanlardan" words that appear in training data is divided by the total number of words in all tweets.

The above formulas apply in the positive case. The positive or negative result from both probability results allows us to estimate the sentiment of the tweet written on this subject. We will classify the tweet as negative because "qorxuram" appears more often on negative tweets.

3.2 Support vector machine

Support Vector Machines are powerful method for building a classifier. This method is based on the concept of decision plans that define decision boundaries. The purpose of this algorithm is to find a decision hyperplane that classifies data points in N-dimensional space (N - number of properties) in a meaningful way. To separate two classes of data points, there are many possible hyperplanes to choose from. Our goal is to find a hyperplane with maximum separation, ie the maximum distance between the data points of both classes. The separation is represented by the vector w.

The optimal hyperplane can then be defined as:

$$wx^T + b = 0,$$

where w is the weight vector, x is the input feature vector, and b is the bias.

We use the SVM software with a linear kernel. Our input data is a set of vectors. Each entry in the vector corresponds to the presence a feature. We use tree types of vectors: Count Vector, TF-IDF and N-Gram.

3.3 Maximum Entropy

Maximum Entropy is a widely used probability distribution prediction technique in the field of natural language processing. This learning method is not different from other learning techniques, in which the outputs of the model are based on the given training data set. Unlike Naive Bayes, Maximum Entropy makes no assumptions about independence about the formation of words. However, it is more expensive in terms of calculation. It is a machine learning method based on experimental data.

Mathematical model of a entropy clustering can be expressed as follows:

$$P_{ME}(c|t, w = \frac{exp\,[\sum_i w_i f_i(c,d)]}{\sum_x exp\,[\sum_i w_i f_i(c,d)]}).$$

In this formula, c is the class, t is the tweet, and w is a weight vector. The weight vectors decide the significance of a feature in classification. A higher weight means that the feature is a strong indicator for the class (Go et al., 2009).

4 Twitter data and processing methods

The data analysis process in this study consists of the following steps:

A. Data collection

The first step in data analysis is to collect data. This is known as "data mining". Data collection is not as simple as it seems at first glance. Data can come from anywhere. The data used in the sentiment polarity classification carried out within the scope of this research study was acquired from Twitter social network. Twitter is a gold data mine. Unlike other social

platforms, almost every user's tweets are public and impressive. If you need large amounts of data to do analytical work, Twitter becomes very important. In addition, Twitter data is quite prominent when compared to other social media data. Since it is easier to obtain data from this microblogging service, twitter social network is generally preferred in the studies. The main reasons for this preference are (Pak et al., 2010):

- Twitter's API allows complex queries, such as pulling every tweet on a specific topic within the last twenty minutes, or pulling a specific user's retweeted tweets.
- Using spatial location information, it provides easy access to data at a specific location.
- This microblogging platform is used by different people to explain their views on different topics, so it is a valuable resource where people share their ideas.
- Twitter contains many text messages, and the number of these messages is increasing day by day. Therefore, the collected corpus can be optionally increased.
- Twitter's audiences range from regular users to celebrities, company representatives, politicians, and even presidents. Therefore, it is possible to collect written messages from users from different social and interest groups.
- Twitter's audience is represented by users from many countries. Although there are more users in the United States, this social network is also widely used in different countries. Therefore, it is possible to collect data in different languages

Twitter is a data mine that creates new challenges in data analysis and has many unique features compared to other fields. Since the data to be analyzed are Twitter messages sent in Azerbaijani language, tweets written in this language were taken from social network. Since there is no Azerbaijani language on Twitter social network, the language of the tweets written is determined automatically in Turkish. Therefore, geographic location information was used to receive tweets written in Azerbaijani language. The point consisting of X and Y coordinates taken randomly from the capital Baku was accepted as the starting point. Twitter data were taken daily from the settled area within a radius of 300 km from this point. Russian, English and Turkish languages are actively used among the Twitter users in Azerbaijani, so in addition to using geographic location when querying, Turkish is also specified as the language parameter. Therefore, tweets written in Russian and English languages are automatically extracted from the incoming data set. Finally, to remove tweets written in different languages in the dataset, the location information used by users who wrote this tweet is checked. If these locations include locations outside the Azerbaijani cities, these tweets are also removed from the dataset.

B. Data cleaning

To apply a classifier on the data, it is important to first process or clear the raw data. The task of preprocessing includes: Hashtag, twitter notations (@, RT), expressions, URLs and stop word (ineffective) removal, identification of slang words and compression of long words. Some of the preprocessing steps performed are described below.

1. *Identifying and removing pointer (usernames and hashtags)*

Twitter allows users to include hashtag (#) and user references (@) in their messages. In tweets, one user can point to other users by using the "@" pointer in front of the username. These pointers have been replaced with the special word "USER_MENTION". Users also tag tweets into a hashtag category using "#". Words and characters marked with these markers are removed from tweets.

2. *Detection and removal of URL*

Many tweets contain URLs that share more than content that can be rendered due to limited character restrictions. The content in the URL can provide additional information about the

feeling the user is trying to express, but it can be very expensive to crawl URLs based on their content. During the preprocessing, all contents associated with the URL have been replaced with the word "URL".

3. *Detection and removal of retweets*

In order to prevent the data from being reprocessed and to avoid duplicate data in the training data, it is checked what "IsRetweet" returns before writing the tweets received using the "Tweetinvi" library. If this property returns "true", no tweets are added to the data set.

4. *Detection and removal of punctuation marks*

It is common for users to use excessive punctuation in a microblogging environment to avoid appropriate grammar and to convey emotions more easily. Punctuation can give an idea about the polarity of the message. For example, exclamation marks are often used to express strong emphasis on polar messages. In this step, irrelevant punctuation has been removed from tweets to avoid unnecessary features.

5. *Detection and removal of numbers*

Numbers are not related to the analysis of text data, so the numbers have been removed from the collected tweets. But if these numbers are used as independent numbers, they are removed.

6. *Character change*

Users use some character combinations in their tweets that add positive or negative meaning to the text. These character combinations are replaced with the specific word equivalent

7. *Convert to lower case*

At last, tweets are converted to lower case. The appropriate equivalents of the Turkish characters are written in the conversion process.

8. Emoji change

Emotions containing sentiment in the tweets were replaced with unicode equivalent. The unicode list of these emojis is kept in a table in the database (Fig. 1). The received tweets were checked character by character and the emoji used were replaced with unicode equivalent. For example, the tweet "Reyiniz ucun tesekkur edirik ☺ (thank you for your comment ☺)" has been converted to the "Reyiniz ucun tesekkur edirik ☺" form before coming to the analysis stage.

	ID	Sequence	EmojiName	Value	Name	SearchTerms	SortOrder	SubGroup
1	71	1F600	GrinningFace	70	grinning face	grinning, face	0	face-positive
2	72	1F601	GrinningFaceWithSmilingEyes	80	grinning face with smiling eyes	grinning, face, smiling, eyes	1	face-positive
3	73	1F602	FaceWithTearsOfJoy	90	face tears joy	face, tears, joy	2	face-positive
4	74	1F923	RollingOnTheFloorLaughing	100	rolling on the floor laughing	rolling, floor, laughing	3	face-positive
5	75	1F603	SmilingFaceWithOpenMouth	60	smiling face with open mouth	smiling, face, open, mouth	4	face-positive
6	76	1F604	SmilingFaceWithOpenMouthAndSmilingEyes	70	smiling face with open mouth & smiling eyes	smiling, face, open, mouth, eyes	5	face-positive
7	77	1F605	SmilingFaceWithOpenMouthAndColdSweat	80	smiling face with open mouth & cold sweat	smiling, face, open, mouth, cold, sweat	6	face-positive
8	78	1F606	SmilingFaceWithOpenMouthAndClosedEyes	90	smiling face with open mouth & closed eyes	smiling, face, open, mouth, closed, eyes	7	face-positive
9	79	1F609	WinkingFace	60	winking face	winking, face	8	face-positive
10	80	1F60A	SmilingFaceWithSmilingEyes	70	smiling face with smiling eyes	smiling, face, eyes	9	face-positive
11	81	1F60B	FaceSavouringDeliciousFood	80	face savouring delicious food	face, savouring, delicious, food	10	face-positive
12	82	1F60E	SmilingFaceWithSunglasses	80	smiling face with sunglasses	smiling, face, sunglasses	11	face-positive
13	83	1F60D	SmilingFaceWithHeartEyes	90	smiling face with heart-eyes	smiling, face, heart, eyes	12	face-positive
14	84	1F618	FaceBlowingAKiss	80	face blowing a kiss	face, blowing, a, kiss	13	face-positive
15	85	1F617	KissingFace	90	kissing face	kissing, face	14	face-positive
16	86	1F619	KissingFaceWithSmilingEyes	90	kissing face with smiling eyes	kissing, face, smiling, eyes	15	face-positive
17	87	1F61A	KissingFaceWithClosedEyes	100	kissing face with closed eyes	kissing, face, closed, eyes	16	face-positive
18	88	263A FE0F	SmilingFace	50	smiling face	smiling, face	17	face-positive
19	89	1F642	SlightlySmilingFace	60	slightly smiling face	slightly, smiling, face	19	face-positive

Figure 1: Code equivalents of emojis

C. Feature extraction

Items in the classification are represented by their properties. In our case, tweets are represented by words, so words are used as a feature. Machine learning requires a basic process to convert text data into a machine-readable format. This process is called feature extraction and there are different methods for feature extraction. The commonly used method is a bag of words model that accepts each word as a feature. Tweets are collected in this model and converted to a list of unsorted words called vocabulary. The next step is to create a vector representation based on the size of the given words. Examples of this vector representation are the following.

Count Vector is a matrix notation of the dataset in which every row represents a tweet from the corpus, every column represents a term from the corpus, and every cell represents the frequency count of a particular term in a particular tweet

TF-IDF Vectors: This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. TF-IDF score consists of two terms:

- TF: Term Frequency is used to measure how often a term occurs in a document. Since the length of each document is different, it is possible for a term to appear more in long documents than in short documents. Therefore, as a means of normalization, the term frequency is usually divided by the document length (ie, the total number of terms in the document).

$$TF\,(t) = \frac{(Number\ of\ times\ term\ t\ appears\ in\ a\ document)}{(Total\ number\ of\ terms\ in\ the\ document)}.$$

- IDF: Inverse Document Frequency is used to measure how important a term is. All terms are equally important when calculating TF. However, it is known that some terms such as "and", "or", "of" may often appear, but of little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:

$$IDF\,(t) = \log_e\left(\frac{Total\ number\ of\ documents}{Number\ of\ documents\ with\ term\ t\ in\ it}\right).$$

5 Developed system for Azerbaijani language

In this section we present the main components of our approach. we developed two different projects for this, because our approach uses both lexicon and machine learning methods. The first project (SAT) implementation was developed using the C# software language and the other (SATPy) was developed using the Python software language. In the SAT application, which is the first of the projects, data from Twitter is received on a daily basis (Fig. 2)

SAT application consists of 4 layers: data collection, sentiment dictionary creation, data processing, lexicon based learning. The data collection layer was developed to connect to Twitter and retrieve data. In this layer, "Tweetinvi", an open source application, was used to retrieve data. The Tweetinv is a NET C # library that allows developers to easily and reliably interact with Twitter.

The second layer was developed to create sentiment dictionaries, because there is no dictionary that contains enough sentiment words for the Azerbaijani language. To do this, download the "SentiWordNet_3.0.0_20130122" library and convert all the words contained here into the Azerbaijani language using the "Cloud Translation API-> translate.googleapis.com" service to create a dictionary containing new emotion words (Figure 3).

The collected tweets are processed in the data preprocessing layer and then sent to the word-based learning layer for search in each word sentiment and emoji dictionary. If words are found in these dictionaries, the sum of emotion score is calculated. If the total score is greater than zero, the tweet is positive, the total score is less than zero, the tweet is negative, and if the

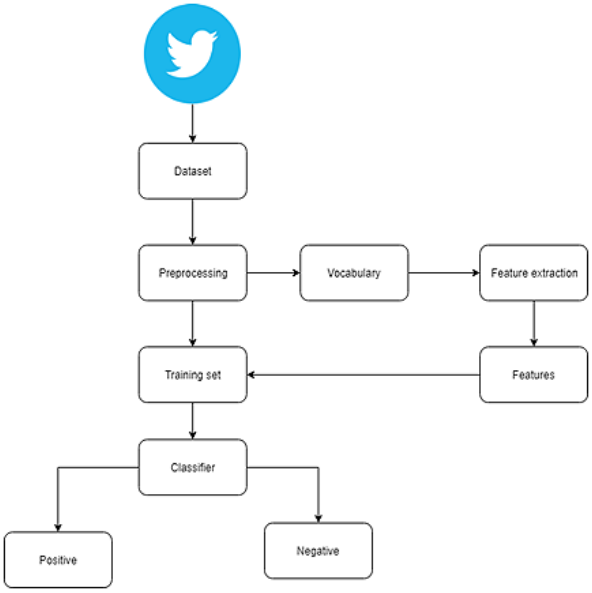

Figure 2: Classification operation

	POS	SentScore	Word_EN	Word_AZ	Word_TR	Polarity
1	a	0.0416666666666667	dying	ölür	ölen	1
2	a	0.202554744525547	absolute	mütləq	kesin	1
3	a	0.25	implicit	örtülü	üstü kapalı	1
4	a	0.0833333333333333	ascetical	sanki	ascetical	1
5	a	0.0340909090909091	greedy	acgöz	açgözlü	1
6	a	0.06	objective	məqsədli	amaç	1
7	v	-0.625	debauch	axmaqlıq	debauch	0
8	v	-0.125	defect	qüsur	kusur	0
9	v	-0.875	humbug	aldatma	riyakârlık	0
10	v	-0.0833333333333333	loll	axmaq	aptal	0
11	v	-0.25	blight	kədər	yıkım	0

Figure 3: Sentiment dictionary

total score is equal to zero, the tweet is neutral. In addition, in the SATPy system developed using python language, machine learning methods were applied using cleared data. For this, positive tweets are read from the database and written to pos_tweets.csv", and negative tweets to "neg_tweets.csv". Afterwards, these tweets were divided into training and test data sets and appropriate machine learning operations were performed.

6 Results

In this work, 238,677 tweets were collected in Azerbaijani language between 01-02-2019 and 12-03-2019 by using geolocation information for analysis. 20,000 tweets of these tweets are manually labeled as positive, negative and neutral. Naive Bayes, Support Vector Machines and Maximum Entropy machine learning methods were applied on labeled tweets.

Table 1: Classification using count vectors

Algorithm	**Accuracy**
Naive Bayes	96.51
Support Vector Machines	95.63
Maximum Entropy	95.63

Naive Bayes method has been shown to give better results in experimental studies using count vectors.

When using the TF-IDF feature extraction vectors of the maximum entropy method, the same count vectors seem to give better results as used.

In the experiments using N-Gram vectors, Naive Bayes and Support Vector Machines are shows same result. However, when compared with other feature extraction methods, the results appear to be very low.

From these results, it would suggest that both "Count vectors" and "TF-IDF vectors" feature extraction are perhaps worthy of further study on this problem.

Table 2: Classification using TF-IDF vectors

Algorithm	**Accuracy**
Naive Bayes	90.39
Support Vector Machines	96.51
Maximum Entropy	96.94

Table 3: Classification using N-Gram vectors

Algorithm	**Accuracy**
Naive Bayes	67.69
Support Vector Machines	67.69
Maximum Entropy	66.81

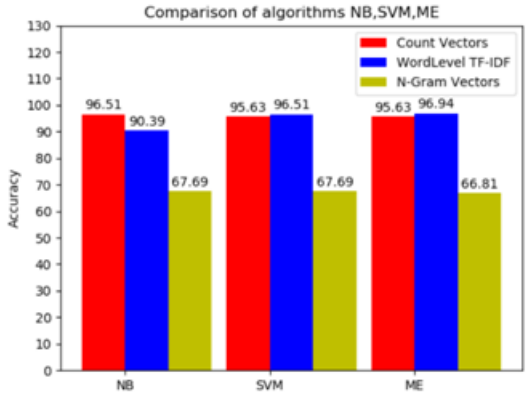

Figure 4: Comparison of algorithms

7 Conclusion

This article is an experimental contribution to the studies in the field of data science and sentiment analysis for the Azerbaijani language. A number of machine learning techniques and word-based learning techniques have been applied to classify sentences based on Twitter data. This study also compares various traditional classification techniques and their accuracy. The main objective is to provide an accurate model for emotional analysis using pre-tagged. We used the collected Twitter data corpus to train a sentiment classifier. Our classifier is able to determine positive, negative and neutral sentiments of tweet. Maximum entropy method yielded better results than the other two methods, Naive bayes and SVM.

References

Aida-zade, K., Rustamov, S., Mustafayev, E. (2013). Sentiment analysis: hybrid approach. *Transactions of Azerbaijan National Academy of Sciences. Informatics and control problems*, *XXXIII*(6), 100-108.

Aida-zade, K.R., Rustamov, S.S., Mustafayev, E.E., Aliyeva, N.T. (2012). Human-computer dialogue understanding hybrid system. *International Symposium on Innovations in Intelligent Systems and Applications*. Trabzon, Turkey.

Aida-zade, K., Rustamov, S., Clements, M.A., Mustafayev, E. (2018). Adaptive neuro-fuzzy inference system for classification of texts. In *Zadeh L., Yager R., Shahbazova S., Reformat M., Kreinovich V. (eds) Recent Developments and the New Direction in Soft-Computing Foundations and Applications*. S*tudies in Fuzziness and Soft Computing*. Springer, *361*(2018), 63-70.

Appel, O., Chiclana, F. & Carter, J. (2015). Main concepts, state of the art and future research questions in sentiment analysis. *Acta Polytechnica Hungarica*, *12*(3), 87-108.

Chory, R.N., Nasrun, M. & Setianingsih, C. (2018, November). Sentiment Analysis on User Satisfaction Level of Mobile Data Services Using Support Vector Machine (SVM) Algorithm. In 2018 *IEEE International Conference on Internet of Things and Intelligence System* (IOTAIS), 194-200. IEEE.

Desai, M. & Mehta, M.A. (2016, April). Techniques for sentiment analysis of Twitter data: A comprehensive survey. In *2016 International Conference on Computing, Communication and Automation* (ICCCA), 149-154, IEEE.

Duwairi, R.M., Marji, R., Sha'ban, N. & Rushaidat, S. (2014, April). Sentiment analysis in arabic tweets. In *2014 5th International Conference on Information and Communication Systems* (ICICS), 1-6, IEEE.

Eliaçik, A.B. & Erdogan, N. (2015). Mikro Bloglardaki Finans Toplulukları için Kullanıcı Ağırlıklandırılmış Duygu Analizi Yöntemi. In *UYMS*.

Go, A., Bhayani, R. & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, *1*(12), 2009.

Hu, M. & Liu, B. (2004, August). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 168-177.

Kumar, A. & Sebastian, T.M. (2012). Sentiment analysis on twitter. *International Journal of Computer Science Issues* (IJCSI), *9*(4), 372.

Lin, J., Li, X. & Jiao, Y. (2010, March). Text Categorization Research Based on Cluster Idea. In 2010 *Second International Workshop on Education Technology and Computer Science* , 1, 483-486, IEEE.

Madhoushi, Z., Hamdan, A.R. & Zainudin, S. (2015, July). Sentiment analysis techniques in recent works. In 2015 *Science and Information Conference* (SAI), 288-291, IEEE.

Mahmood, T., Iqbal, T., Amin, F., Lohanna, W. & Mustafa, A. (2013, December). Mining Twitter big data to predict 2013 Pakistan election winner. In *INMIC*, 49-54, IEEE.

Medhat, W., Hassan, A. & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093-1113.

Pak, A., Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, 1320-1326.

Pang, B., Lee, L. & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, 10, 79-86. *Association for Computational Linguistics.*

Rane, A. & Kumar, A. (2018, July). Sentiment classification system of Twitter data for US airline service analysis. In 2018 *IEEE 42nd Annual Computer Software and Applications Conference* (COMPSAC), 1, 769-773, IEEE.

Rustamov, S., Clements, M. (2013). Sentence-level subjectivity detection using neuro-fuzzy model. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Association for Computational Linguistics*, Atlanta, 108-114.

Rustamov, S., Mustafayev, E., Clements, M.A. (2013). An application of hidden Markov models in subjectivity analysis. In *2013 7th International Conference on Application of Information and Communication Technologies*, IEEE, 64-67.

Rustamov, S., Mustafayev, E. & Clements, M.A. (2013, April). Sentiment analysis using Neuro-Fuzzy and Hidden Markov models of text. In 2013 *Proceedings of IEEE Southeastcon*, 1-6, IEEE.

Rustamov, S. (2018). A hybrid system for subjectivity analysis. *Advances in Fuzzy Systems*, 9.

Rustamov, S.S. (2012). An application of neuro-fuzzy model for text and speech understanding systems. *PCI'2012. The IV International Conference Problems of Cybernetics and Informatics,* Baku, Azerbaijan, I, 213-217.

Rustamov, S.S. (2012). An application of neuro-fuzzy model for text and speech understanding systems. (2012). *The IV International Conference Problems of Cybernetics and Informatics*, Baku, Azerbaijan, I, 213-217.

Rustamov, S.S. (2012). On an understanding system that supports human-computer dialogue. *PCI'2012. The IV International Conference Problems of Cybernetics and Informatics*, Baku, Azerbaijan, I, 217-221.

Rustamov, S.S. (2012). On an understanding system that supports human-computer dialogue. *The IV International Conference Problems of Cybernetics and Informatics*, Baku, Azerbaijan, I, 217-221.

Rustamov, S., Mustafayev, E., Clements, M. A. (2013, April). Sentiment analysis using Neuro-Fuzzy and Hidden Markov models of text. In *2013 Proceedings of IEEE Southeastcon*, 1-6, IEEE.

Saravanan, M., Sundar, D. & Kumaresh, V.S. (2013, December). Probing of geospatial stream data to report disorientation. In 2013 *IEEE Recent Advances in Intelligent Computational Systems* (RAICS), 227-232, IEEE.

Satı, N.U., & Ordin, B. (2018). Application of the polyhedral conic functions method in the text classification and comparative analysis, *Scientific Programming*, Volume 2018, Article ID 5349284, 11 pages.

Shital, A.P., Jeevan, A.P. (2017). Twitter sentiment classification using stanford NLP. *1st International Conference on Intelligent Systems and Information Management* (ICISIM).

Smith, P., (2011). Sentiment analysis: beyond polarity. Doctoral dissertation, Thesis Proposal, School of Computer Science, University of Birmingham, UK.

Suleymanov, U., Rustamov, S. (2018). Automated news categorization using machine learning methods. *IOP Conference Series: Materials Science and Engineering*, 459, 012006.

Suleymanov, U., Rustamov, S., Zulfugarov, M., Orujov, O., Musayev, N., Alizade, A. (2018). Empirical study of online news classification using machine learning approaches. In 2*018 IEEE 12th International Conference on Application of Information and Communication Technologies* (AICT). IEEE, 1-6.

Velioğlu, R., Yıldız, T. & Yıldırım, S. (2018, September). Sentiment analysis using learning approaches over emojis for Turkish Tweets. In 2018 *3rd International Conference on Computer Science and Engineering* (UBMK), 303-307, IEEE.

Yadav, S.K. (2015). Sentiment analysis and classification: a survey. *International Journal of Advance Research in Computer Science and Management Studies*, *3*(3), 113-121.