ISSN 1512-1232

TWITTER AND SENTIMENT ANALYSIS

Zh.G.Gogiashvili, O.M.Namicheishvili, V.A.Prangishvili and N.O.Namicheishvili

Georgian Technical University
o.namicheishvili@gtu.ge, j.gogiashvili@gtu.ge, vprangis@gmail.com, n.namicheishvili@gtu.ge

Summary: The designed framework collects data from tweets and uses natural language processing techniques to extract features. The natural language processing is then applied to classify the sentiment as positive, negative, and neutral. Polarity and partiality are also calculated by the dictionary, which consists of a semantic evaluation of the tweet. It has been observed that natural language processing is a better method for sentiment analysis than traditional methods. There are some limitations in natural language processing, so other machine learning and data mining techniques may be used in the future to address the limitations of these feature vectors and their selection. Future work will focus on a multilingual machine learning algorithm that processes different types of tasks and easily categorizes data into groups and evaluates them based on real-time data opinions.

Keywords: tweet, sentiment, natural language processing, feature extraction, sentiment classification, machine learning, multilingual machine learning algorithms.

Introduction

A person cannot a priori know everything in the world. Often there are situations in life when it is necessary to obtain information or make a choice in an area of knowledge about which an individual knows almost nothing. This is when one has to turn to outside help. If decades ago we used to ask our friends, relatives and acquaintances for traditional advice, now everything has changed. With the rapid development of information and computer technology, and in particular the global web Internet, a worthy alternative for finding the necessary information and help in choosing something has appeared.

It would seem that in the age of modern technology, what could be easier than to send a query to a search engine, and it, in turn, will give answers to all the user's questions. But do such search tools really help to fully satisfy a person's information needs? Because of the enormous amount of different content in the world wide web, which is growing rapidly every day, very often relevant information gets lost among the enormous variety of useless data. In addition, traditional information searches and web searches in particular do not always help in finding third-party opinions to make one's own decision.

At the same time, the last decade is characterized by increasing popularity of various social systems: blogs (e.g., LiveJournal[1], Twitter[2]), forums (a huge number of thematic communities, for example, TripAdvisor[3] - a forum of travelers, Cyberforum[4] - forum programmers), social networks

[1]http://www.livejournal.com/
[2]https://twitter.com/
[3]https://www.tripadvisor.ru/ForumHome
[4]http://www.cyberforum.ru/

ISSN 1512-1232

(e.g., Facebook[5], Instagram[6]), online services, accumulating opinions about a particular object (e.g., Amazon[7]). Every day, users of such resources post a lot of messages, materials, express their opinions about a particular object. On the basis of such comments a person can conclude whether or not to use the service of interest, to buy or not the desired product. At the moment, despite all the usefulness of this approach to monitoring opinions, there are a number of serious drawbacks: difficulties in manual processing of huge volumes of data, finding opinions and their emotional evaluation, bringing the result to a convenient form.

Based on the above, there is a need to create a system for automatic finding and analysis of opinions. Such a task is posed in a discipline that lies at the intersection of information retrieval and computer linguistics - text tonality analysis and opinion mining - sentiment analysis and opinion mining. Sentiment analysis is a system of automatic extraction of emotionally colored vocabulary and opinions from texts in relation to the objects the text is about. Tone is the emotional attitude of the author of the statement, to some object expressed in the text. Opinion usually refers to an emotional evaluation of something.

As the complete name of the subject of sentiment analysis makes clear, the entire discipline can be divided into two large parts. The first is textual tonality analysis, which often involves the task of classifying a corpus of documents based on the tonalities found in them. The second part, opinion extraction, usually aims to isolate all opinions about the objects of interest from the corpus of documents.

The tasks of both blocks of sentiment analysis are relatively recent, so work on them is ongoing. Despite the existence of existing tools and platforms that make it possible to determine not only the tone of messages in social media, but also to identify topics under discussion, analyze opinions about brands, and analyze some other parameters, there is no single accurate algorithm for solving this problem. Consequently, the task of building a system for extracting opinions and analyzing tones is still relevant.

The task of analyzing tones and extracting opinions is quite young - it is a little over a decade old. The rapid development of the Web and the genuine interest in this discipline in particular, and in the field of natural language processing in general, has prompted the scientific community to create a large number of works and articles related to the topic of sentiment analysis.

The terms tonality and opinion were introduced in [11, 17]. The first works of the researchers in this field were characterized by a narrow focus and were exclusively of applied nature. Thus, in [11] methods for obtaining "reputations" - the numerical values of words used in the vicinity of a meaningful word (product mentions on the Internet) were presented. In this paper, the determination of tones was based on the work with dictionaries.

In [13], one of the first comprehensive reviews of the entire field of opinion research is carried out. It touches on the topics of tone detection, opinion selection, the complexities involved in analyzing comparative sentences, and finding spam in opinions.

[5]https://www.facebook.com/
[6]https://www.instagram.com/
[7]https://www.amazon.com/

ISSN 1512-1232

Studies [7, 17, 21] consider classification tasks at the document level, where reviews of services are broken down into negative or positive ones, reflecting the authors' opinions about these services. In [8, 9], the key attention is paid to the extraction of objects and their characteristics from unstructured documents.

1. Communicational system of twitter

Let's look in more detail at Twitter's communication system. Twitter is a system that allows users to send short text notes (up to 140 characters) using the web interface, SMS, instant messaging services or third-party client programs. A user can add and read messages on Twitter by going to www.twitter.com. To understand what Twitter is for, we need to know the features of Twitter and how it differs from other social networks. The features of Twitter are as follows:

1. Information on Twitter about what has happened, what is happening and what is planned spreads faster than on any Internet site or social network. According to some data, the speed of information dissemination today: Twitter ~ 5 minutes, Ribbons ~ 1 hour, Radio ~ 30 minutes, TV ~ 2 hours.

2. Twitter messages are available to all Twitter users, which is (as of January 1, 2011) more than 200 million users! And that number is growing at an unprecedented rate.

3. Twitter is trendy. In promotion of their companies, in public relations, twitter is used only by those companies, which follow modern tendencies and go "with the times".

So if you have Twitter, you do not just follow some fashion, you can make your company, your product or your message known instantly to millions of users.

How do you do it? How does Twitter work and how does it work? Any registered Twitter user can send messages of no more than 140 characters. The messages can contain any information (opinions, news, ideas, events), as well as links to web pages (articles, news, anecdotes, useful information). Links to videos and images (photos, pictures, jokes, music) are particularly popular. An important parameter of popularity of a Twitter user is: the number of followers (those who follow your tweet) and the number of retweets (citations, repetitions) of his tweet. If your message is quoted, it means that it is something interesting, something that should become known to others. The retweet system is a way to make a post known to friends (followers) and friends of those friends. That is, by adding one message, a multi-million audience can make it known to others if it seems important or interesting to them. In fact, Twitter allows you to promote a company for free among a large, modern and mobile audience.

Nowadays the functioning of Internet discourse as a culturally and socially defined communicative activity takes place through the production, storage and transmission of information about the processes of the surrounding reality, which is projected into the mass consciousness of society. It is the social network Twitter that is actively developing a platform within which the way of organizing Internet discourse is being transformed. The number of registered users exceeds 3 million people, and the monthly audience is more than 4 million people.

Social media (Twitter, Facebook, LinkedIn) is probably the most popular free public forum available for the general public to express their thoughts on a variety of subjects. Millions of posts every day - there is a wealth of information hidden there. In particular, Twitter is widely used by companies and ordinary people to describe the state of affairs, promote products or services. Twitter

ISSN 1512-1232

is also an excellent source of data for intelligent analysis of texts: from the logic of behavior, events, the tone of statements to predictions of trends on the stock market. There lies a huge reservoir of information for the intellectual and contextual analysis of texts. The topics of the messages are usually:

1. An ongoing promotion or company news of interest to many people, an interesting video or photo related to a certain industry.
2. Links to useful materials on the industry.
3. News or links to news about an industry you are interested in.
4. Opinions of your company on any issues being discussed in the country.
5. .Advice on matters in which your employees are competent.

2. Exploring the use of sentiment analysis on Twitter

Of interest are the theoretical aspects of the analysis of tones and the selection of opinions that are reflected in social networks and, in particular, in Twitter.

The task of tone analysis means finding lexical tones (lexical sentiments, sentiment words) in a corpus of documents - emotional components expressed at the lexeme level, for the purpose of further classification of documents of this corpus with the help of the found sentiment words. A lexeme is defined as an instance of a sequence of characters in a particular document, combined into a semantic unit for processing. This task is also called the task of document polarity classification, i.e. it is determined whether the opinion expressed in the document is positive or negative (in the simplest case).

The tonalities revealed in the corpus can be classified in different ways, depending on the model chosen. Quite often a one-dimensional emotive space with the polarities "positive" or "negative" is used. However, more complex approaches are sometimes used quitesuccessfully.

1) Classification on a binary scale [17, 21].

The most common approach, which often uses two classes of evaluation: positive and negative. Despite the apparent simplicity of this approach, it is not always possible to unambiguously determine which class a document can be assigned to: an evaluative text may contain signs of both positive and negative evaluation.

2) Classification according to a multipole scale [16, 18].

The most obvious way to complicate the previous approach is to increase the number of classes. The polarity grading now has more than two items. The first works with the corresponding approach were aimed at classifying reviews/reviews on a multi-point scale.

3) Scaling systems [20].

Another approach to determining tonalities is the use of scaling systems, whereby words sentiments are assigned numbers on some discrete scale, such as -5 to +5 (from sharply negative to sharply positive). Next, the text is analyzed using natural language processing algorithms, and then the objects extracted from the text are examined in order to understand the meaning of these words.

4) Subjectivity/objectivity [19].

ISSN 1512-1232

Another research direction is the identification of subjectivity/objectivity. In this task, this text belongs to one of two classes: subjective or objective. This approach goes towards complicating the methodology of the usual polarity classification: the subjectivity of words and phrases can depend on the context, and an objective document can contain subjective sentences.

The question of the currently known tonality analysis algorithms also logically arises.

Tone analysis can be divided into 2 separate categories:

1) manual (tone analysis by assessors);
2) automatedtone analysis.

The difference between the two lies in the accuracy and efficiency of the analysis. The expert, of course, processes the input data much more correctly, but cannot compete with a computer in the volume and speed of processed data arrays.

The following algorithms are often used for automated tone analysis:

1) Based on rules [14].

The approach is to generate the rules on the basis of which the tone of the text will be determined. To do this, the text is divided into words or sequences of words. Then the data obtained are used to identify frequently used patterns, which are assigned a positive or negative score.

2) Using sentiment word dictionaries.See for example:

http://ptrckprry.com/course/ssd/data/positive-words.txt

http://ptrckprry.com/course/ssd/data/negative-words.txt

Often, along with the previous approach, work with sentiment word dictionaries is used. According to the lexical tones found in the text, it can be rated on a scale containing the amount of positive and negative vocabulary. The simplest estimate is the arithmetic mean of all the polarity values of sentiment words.

3) Machine learning without a teacher [21].

This approach is based on the idea that the terms that are most frequently found in this text, and at the same time are present in a small number of texts of the whole collection, have the greatest weight in the text. By identifying these terms and determining their tonality, we can infer the tonality of the entire text.

4) Machine learning with a teacher [10].

This approach requires a learning collection of texts marked within the emotive space, on the basis of which a statistical or probabilistic classifier is built.

5) Based on graph theory models [2].

Based on the assumption that not all words in the text are equivalent, a graph is constructed. In performing this procedure, we find the vertices that have more weight, and thus contribute the most to the determination of the tone of the text. After that we classify the found words on the basis of tone dictionaries.

ISSN 1512-1232

6) Hybrid method.

This method combines all or several approaches considered above, and consists in the application of classifiers on their basis in a certain sequence.

In practice, it is very important to assess how well the correctness and quality of systems for analyzing the tonality of texts agree with the expert's opinion on the emotional coloring of the data presented. In disciplines closely related to information retrieval, the metrics of completeness (recall) and accuracy (precision) are traditionally used as such assessments [15]. Possible evaluations by the system and the expert are shown in Table 1.

Table 1.

Possible evaluations by the system and the expert

		Expert evaluation	
		Positive	Negative
System Evaluation	Positive	TruePositive	False Negative
	Negative	False Positive	True Negative

Then the metrics will be calculated by the following formulas:

$$\text{Accuracy} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Completeness} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

When classifying documents by polarity for each class responsible for a separate tone, the metrics can be calculated as follows:

$$\text{Accuracy} = \frac{\text{Total number of documents assigned to this class with this tone}}{\text{Total number of documents assigned to this class}}$$

$$Completeness = \frac{\text{Total number of documents assigned to this class with this tone}}{\text{Total number of documents with this tone}}$$

The question of finding opinions-emotional judgments about an entity or aspect of it, expressed by some subject-is of utmost importance.

ISSN 1512-1232

For the task of opinion extraction, as is clear from the name, the main goal is to find all emotionally colored opinions about something in the text. In general, an opinion can be expressed about any subject: a product, a service, a person, an organization, an event, etc. In order to distinguish the entity that the text is about, we will use the term *object*.

Each object has a set of *components* (or *parts*) and a set of *attributes* (or *properties*). Each component may contain its own personal subcomponents and its own set of attributes. Thus, each object is a hierarchical structure based on the "*consists of*" (*part-of*) relation.

Formal definition of an object: an object o is an entity that can represent a single object, a person, an organization, an event. An object can be defined by a pair **o**: **(*T*, A)**, where ***T*** is responsible for the hierarchy of components, ***A*** is the set of attributes of the object o.

In practice, however, this definition is usually simplified due to the complexity of natural language processing tasks. Therefore, the hierarchy is intentionally made flat: the term ***feature***(characteristic) is used for both components and attributes.

The opinion holder is the subject who expresses an opinion. Considering the specifics of social systems, we can conclude that most often the authors of opinions are the authors of the posts or messages themselves.

An opinion about a characteristic**f** is a positive or negative view, evaluation, or emotion about a characteristic f expressed by the author of the opinion.

The polarity of an opinion (opinion orientation) about a characteristic **f** indicates whether the opinion is positive, negative, or neutral.

Thus, a model of the object is constructed: the object o is expressed by a finite set of characteristics: $\boldsymbol{F = \{f_1, f_2, \dots, f_n\}}$, which includes the object itself as a special characteristic. Each characteristic $\boldsymbol{f_i}$ can be expressed by a finite number of words or phrases $\boldsymbol{W_i = \{w_1, w_2, \dots, w_m\}}$, which are ***synonyms*** of this characteristic.

An opinion can refer to one of the following two types [12]:

1) *Direct opinion:* formally, it is a tuple of 5 elements $(o_j, f_{jk}, oo_{ijkl}, h_i, t_l)$, where o_j is the object, $f_{jk}$is the object characteristico_j,$oo_{ijkl}$is the polarity of opinion about the object characteristico_j,$h_i$is the opinion author, t_l is the time mark when the opinion was expressed by the author. The polarity of opinion can be not only positive, negative, or neutral, but can also represent a more graded scale to account for varying degrees of emotionality.

2) *Comparison:* a comparison expresses a relation of similarity or difference between two or more objects, and/or an expression of the opinion's author's preferences based on a comparison of common characteristics of objects. This type of opinion is usually characterized by the use of comparative or superlative degrees of adjectives.

Depending on the problem to be solved, sentiment analysis can be conducted at different structural levels [12]:

1) *At the level of the document:* at this level, the entire document is classified as positive, negative or neutral (or according to the polarity scale chosen).

ISSN 1512-1232

2) *At the sentence level:* each sentence of the text is classified as positive, negative, or neutral (or according to the polarity scale chosen). Problem solving at this level is characteristic of comparative sentences.

3) *At the level of characteristics*: finding all the opinions expressed about an object, or its characteristics; determining the tones of opinions. In essence, the task is identical to the task of extracting opinions.

Thus, the tasks of sentiment analysis include:

- classification of documents based on opinions,

- classification of proposals on the basis of subjectivity and opinions,

- aspect-based opinion analysis,

- abstracting opinions based on aspects,

- creating a dictionary of opinions,

- searching for comparisons,

- searching for spam in reviews,

- analyzing the usefulness of reviews,

- relationship search,

- link recognition,

- synonym extraction,

- other information extraction tasks.

Tone analysis finds its practical application in a multitude of areas:

- sales and marketing - based on the data of social systems monitoring, conclusions are made about the popularity of a particular product, finding current trends among customers;

- politics - analyzing data on users' political positions, predicting election results;

- Finance - forecasting markets based on news, blogs and other social media;

- security - monitoring public sentiments, suppressing planned illegal actions;

- Sociology - extracting the social data of the users of interest: political, religious views;

- other areas.

It is interesting, for example, to create a classifier **to see if this or that message is populist or not?**

All political science definitions of populism are based in one way or another on the dichotomous opposition of elites and people (for example, Princeton University's Wordnet gives the definition: populism - the political doctrine that supports the rights and powers of the common people in their struggle with the privileged elite). On the everyday level, populism is the cynical, demagogic statements of politicians in order to win or retain the support of the masses. Populism has no integral ideological or political doctrine; it would be a simplification to equate any manifestations of

ISSN 1512-1232

demagogy and cynicism in politics with populism. It is more correct to define populism as a qualitative characteristic of political doctrines, parties and movements for which the opposition of the elites and the masses is central or one of the most important items on the agenda, and as a method and style of mobilizing mass support for such forces and doctrines. Populist movements and doctrines are a synthesis of such opposition with other political programs, to which populism gives additional mobilizing power and poignancy. In its essence, populism is an institutional problem. Its main message is about the quality of the representation of interests in politics. Adequate political representation is usually understood as the most accurate consideration of the various opinions and interests existing in society, power politics (Farrell D. Electoral Systems: A Comparative Introduction. London: Palgrave Macmillan, 2011. P. 10-11.). From this perspective, populism is a struggle against the inadequate representation of the interests of the general population by elites, but at the same time, the main institutional feature of populism is the lack of respect for pluralism. Appealing to the masses, to the majority, it seeks to fill the entire political arena with this narrative, to free it from all mediating institutions (primarily parties) and procedures. This is the summary of an overview of theoretical knowledge of populism in the work of N. Urbinati (Urbinati N. Democracy Disfigured: Opinion, Truth and the People. Harvard University Press, 2014. P. 131-145.). Thus, there are deep institutional contradictions in the very nature of populism. First, it is by definition "anti-elitist," but its preachers in the political space are a certain part of the political elite, which becomes the beneficiary of popular support for anti-elitist ideas. Second, by downplaying the role of institutions, it threatens the integrity of the entire political system, especially the component that ensures the responsibility of power and the protection of minority interests. If pluralism is undeveloped in the political system, populism effectively monopolizes power.

Pang and Lee (2008, [1]) saw possible applications of sentiment analysis in many areas. One of these areas was politics, among others. Needless to say, they were right. In recent years sentiment analysis has been applied from various sides of research in the political sphere.

Williams and Gulati (2008, [3]), for example, analyzed the influence of Facebook during the 2008 U.S. presidential election. What emerges from this study is that the voter support that a White House candidate receives on Facebook strongly reflects the degree of the latter's success during his campaign. There is even a correlation between the online support a candidate receives in a particular state of the country and his or her voting record. However, a high degree of online support in no way compensates for weak or absent policies. Cornfield (2008, June 4) explains the Obama campaign as follows (*Yes, it did make a difference*. Media &Politics. Consulté le 19 février, 2018 sur http://takingnote.tcf.org/2008/06/yes-it-did-make.html) :

«Не будь Интернета, не было бы и Обамы, потому что ему удалось собрать большую сумму денег для финансирования своей предвыборной кампанию посредством онлайн-активности» (цитируется по Williams and Gulati, 2008, [3]).

In 2008, the candidate who generated the most online activity and gathered the most supporters online was Obama. The latter also ended up winning the U.S. election with 52.9% of the vote, 7.2% more than his opponent. This event is considered one of the most surprising results of the American elections.

A year later, Williams and Gulati (2009, [3]) investigated the use of social media, and Facebook in particular, during the 2006-2008 US congressional elections. Already during this period, the two

ISSN 1512-1232

researchers noticed that more and more people were using social networks to spread their political opinion and that more than 70% of the House candidates had a Facebook page. The main purpose of these pages was to win the votes of the youngest voters.

Van Hee (2013,[4]) in her corpus study tries, among other things, to find out whether a link can be established between political tweets and election results in Flanders. Similarly, Tumasjan et al. (2010, [5]) analyzed political tweets during the 2009 German federal elections and concluded that Twitter does serve as a platform for political debate and that the same social network does reflect the political current that prevails in the material world. Thus, they describe that political tweets are not only used to express political opinion, but also to discuss various political opinions with other users. In addition, the researchers found a correlation between the total number of tweets (100,000) and the election result. The political parties that appeared together in the tweets also represented the coalitions that existed at the time. However, it is important to note that a small number of users (4% in their analysis) were the authors of about 40% of the tweets.

In 2012, Boullier and Lochard [6] also pointed out in their book that social media monitoring (active monitoring of certain topics in social networks to determine the general opinion of the latter) can be applied in the field of politics: "a step in this direction could be to study cases of quotes, arguments, criticism caused by the words of a candidate, elected representative, party, etc.

Nowadays, there are already a fairly large number of ready-made systems for analyzing tones and finding opinions. Here are some of them.

1) Stanford NLP [26] is an open demo model from Stanford University that allows determining the tone for movie reviews. The system is based on the use of recursive neural networks. It supportstexts in English only.
2) Sentiment140 [25] is a solution developed by Stanford graduates. It positions itself as a tone analyzer for Twitter microblogging system. It allows the user to get a selection of positive, negative or neutral micro-messages in response to his request. It visualizes the corresponding result using infographics. The service works only with English and Spanish languages.
3) 30dB [28] - free platform. Similarly to the previous service, it accepts a request for input and outputs emotional opinions about the received topic. As data for analysis it uses not only Twitter, but also Facebook, Google+. As an additional option allows you to compare the emotional component of the two entered topics at once. Supports the English language only.
4) VAAL [23] is a Russian development started in 1992. The system allows you to predict the effect of the unconscious impact of texts on the mass audience, analyze texts in terms of such impact, identify the personal and psychological qualities of the authors of the text, perform emotional and lexical and content analysis of texts, produce automatic categorization of the text.

ISSN 1512-1232

3. Automation and Robotics in Sentiment Analysis

Automation and robotics are two elements that are no longer terms unknown to the general public today. Whether in computer science or artificial intelligence, scientists are constantly finding new methods to automate long, time-consuming and even boring procedures. Within automated language processing, the scientific community has been working for years to develop programs to improve the speed and consistency of analysis. Unlike humans, programs work much faster and more methodically in this area. This is where machine learning comes into play.

The subdivision of artificial intelligence under machine learning refers to the methods and actions that researchers need to develop an automatic system that, by analyzing data, learns itself. In contrast to "rule-based learning," systems do not contain rule sets but rather algorithms, "sets of explicitly programmed instructions used by systems to compute or solve problems"

(Raphaël, 2017, https://www.supinfo.com/articles/single/6041-machine-learning-introduction-apprentissage-automatique). These Thus, learning algorithms allow systems to train with the data provided in order to evolve and improve.

There are three main methods of developing an automatic system. The first method, which is also the most common, is called supervised learning. This technique requires the most human interaction. Indeed, this method implies that the system receives data already analyzed, annotated, and classified by experts as an example. Thus, this data allows the system to develop its own algorithm so that it can analyze and label the raw data on its own. This is why this type of learning falls into the category of inductive machine learning. Currently, supervised machine learning has many applications, such as automatic language processing and handwriting recognition (Kotsiantis, S.B., Zaharakis, I., &Pintelas, P., 2007, Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160, 3-24).

For the second method, called teacherless learning, the developer provides only the raw data to the system. In this case, it is the same system which, thanks to its algorithm, will itself look for different possibilities to classify and analyze the data. This method proves to be interesting as it allows finding new ways to classify data (Marrone, R., and Möller, R., 2011, Foundations of Machine Learning and Data Mining [class handout]. Hamburg : the Technical University of Hamburg, ES42.).

Finally, reinforcement learning is the most sophisticated learning method, which is often used in strategy games. In this method, an automatic system is exposed to a so-called "environment," such as a maze. The system then tries to perform the actions itself, to which it assigns a value (weight). Finally, after analyzing all the possibilities, the system will choose the best action to take, depending on the situation it is in (the best possible reward) (Marrone and Möller, 2011; Kotsiantis, 2007). In the example given by these authors, the system took three actions. The first attempt took it 120 seconds, the second 50 seconds, and the last 85 seconds. In this case, the system will assign the highest value to the second attempt, since the last one was the best.

According to Kotsiantis (2007), it is not important to classify automatic learning systems according to their effectiveness. What is really important is to find out which systems perform better and in which areas.

ISSN 1512-1232

4.Supervised machine learning process

To develop a supervised machine learning system, there are a few standard steps to follow.

It is logical that such systems are designed to achieve a certain goal. The main topic of this thesis will be used as an example: we want to analyze the feelings that Georgians are expressing on Twitter about the new President Salome Zurabishvili. Although these steps can be done manually, the usual practice is that they are done automatically using specialized programs.

To do this, the first step is to collect data about Salom Zurabishvili. Therefore, tweets referring to Salom Zurabishvili will be collected manually or automatically by software/software. During the collection, the researcher will also annotate the data and label it. For example, tweet №X has a negative feeling, talks about the housing tax and does not contain irony.

Then comes preprocessing. This step involves processing the data so that the system can easily analyze it. In other words, the data is pre-filtered to improve its quality. To achieve this goal, there are a large number of methods such as stop-word removal (extremely common words), tokenization (to separate data into individual words), lemmatization (to visualize words in their canonical forms), feature extraction, etc. (Vijayarani, S., Ilamathi, M. J., & Nithya, M., 2015, Preprocessing techniques for text mining-an overview. International Journal of Computer Science & Communication Networks, 5(1), 7-16.).

At this stage, the scientist must also divide the collected data into three sets: an annotated training set with labels to train the system, a validation set to validate its performance, and a test set to evaluate the results. Although the ratio for these three categories varies from study to study, Marrone, R. and Möller, R. (2011, Foundations of Machine Learning and Data Mining [class handout]. Hamburg : the Technical University of Hamburg, ES42) recommends a distribution of 50% - 25% - 25% respectively.

The next step, a critical one according to experts, involves the choice of a learning algorithm. A critical step because it is the algorithm that will completely determine the way the analysis system works. There are several algorithms used in machine learning systems.

Once the algorithm has been selected, the last step is to train the automatic system. The training set is then made available to the system as input data. Since this set includes annotations and labels, the system receives the response data and then uses this set to try to classify the new raw data it encounters. To see if the system gives satisfactory results, a set of scores containing the raw data is offered for analysis. If the results are good, the system finally receives the test set as the final score. If, on the other hand, the results are unsatisfactory, the researcher will have to make changes to the previous steps.

5.Supervised (Controlled) machine learning algorithms

Algorithms can be seen as the heart of machine learning systems. Without these instruction sets, systems cannot learn on their own. Regarding supervised machine learning, several algorithms can be implemented in practice and are divided into two groups. On the one hand, these are logic-based algorithms such as decision trees, and on the other hand, statistical algorithms such as neural networks and SVMs (Kotsiantis 2007).

5.1 Decision Trees

As the name implies, decision trees are algorithms that represent their decisions in the form of a tree. The decisions made by the system are read from top to bottom and represent a hierarchy of decisions. The top of the tree is the root node. Each node represents a characteristic, and the

ISSN 1512-1232

resulting branches contain a leaf that represents a possible solution. The leaf selected from the last node of the tree contains the final solution to the system. Decision trees are advantageous because they are fast and, above all, readable. However, these algorithms run the risk of generating too many branches, which can slow down the speed of computation. In addition, because these algorithms only deal with one node at a time, it is difficult for them to deal with problems that require multiple features to be considered simultaneously (Kotsiantis 2007).

5.2 Artificial neural networks

Artificial neural networks are complex algorithms that attempt to mimic the workings of the human brain. As Haykin, S. S. (1999,. Neural networks: a comprehensive foundation, 2nd ed., Upper Saddle River, N.J: Prentice Hall), the human brain works with flexible neurons that are constantly learning and adapting to their environment. Similarly, these algorithms contain artificial neurons that try to work the same way. To do this, a neural network contains billions of interconnected neurons that work in parallel. This allows data to be processed on multiple aspects simultaneously, which is not possible with a decision tree. The strength of the connection between neurons is called "synaptic weight." After each action, the system stores its experience in its neurons in an attempt to improve its future results. In other words, it learns from the environment in which it is immersed. Each experience will lead to a change in "synaptic weights." Thus, the weight of the connections between different neurons will determine how the system will function (Haykin, 1999). These algorithms also have the spirit of generalization. They can generate values for unseen data based on similar previous experiences.

It is important to note that a neural network cannot work alone. Indeed, to achieve convincing results, the system must have multiple neural networks so that they can work together. Each new problem the system encounters is divided into several smaller problems, which are then sent to different parts of the network. Neural networks are used in many disciplines, such as mathematics, computer science, and word processing, without showing the best results that can only be achieved in so-called deep neural networks.

5.3 Naïve Bayesian classification

Naive Bayesian classification forms a much simpler linear classification system than neural networks. Indeed, this statistical algorithm operates on the basis of probabilities and hypotheses. Although the term "naive" may seem strange, it refers to the thinking of the system. This algorithm assumes that a characteristic belonging to a class forms no relation with other possible characteristics (Kotsiantis, S. B., Zaharakis, I., &Pintelas, P. 2007, Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160, 3-24.; Leung, K., 2007, 'Naive Bayesian Classifier', Technical Report, Department of Computer Science / Finance and Risk Engineering, Polytechnic University, Brooklyn, New York, USA.). This explanation seems rather abstract, and so we accompany it with an example.

A bird has great related characteristics: it has wings, feathers, a beak, and so on. If we ask a person what animal has wings and a beak, he will immediately recall a bird without thinking of other animals. If a naive Bayesian classification system receives an image of a bird as input, it first identifies these characteristics. However, the algorithm does not establish a relationship between these items, but rather analyzes them one by one. After calculating the probabilities of each characteristic, the system will try to figure out which animal is the most sensitive to have wings, feathers, a beak, two eyes, and formulate its hypothesis: according to the calculated statistics, this image must be a bird. This rather simplistic operation allows the system to look quickly at the input.

ISSN 1512-1232

This is the reason why it learns quickly. Several comparative studies have shown that this algorithm generates results comparable to those of decision trees. Nevertheless, speed remains the main advantage.

5.4 The k-nearest neighbor algorithm

The nearest neighbor algorithm is a simple instance-based learning algorithm. Belonging to the family of statistical methods, the latter also behaves as a lazy algorithm, which means that it does not learn the training data but remembers them (Kotsiantis, 2007). This means that the learning phase of this algorithm is fast. However, the system must have access to its memory for each computation, and the classification speed is quite low compared to other methods (Cunningham, P., and Delany, S. J., 2007,. k-Nearest neighbour classifiers. Multiple Classifier Systems, 34, 1-17.; Kotsiantis, 2007). Unlike the other algorithms mentioned above, the k-nearest neighbor algorithm works quite simply. The memory of such an algorithm resembles a space full of points, each point representing a set of known data. When the system receives a new entry for classification, the latter first converts it into a point and inserts it into its memory. The algorithm will then search for k points closest to the new point. After determining these nearest points, the system will look for which classification appears most often around the new point. It will then be assigned a majority class. The new classified data, in turn, will be stored in system memory to serve as a reference for new entries (Cunningham and Delany, 2007). Since the system works with majority voting, it is desirable to take an odd number k so as not to risk the equality of the number of classes. Many recommendation systems use the nearest neighbor algorithm.

5.5 Support vector machines

Reference vector machines, abbreviated as SVMs (Spport-Vector Machines), basically form a linear system which, using kernel techniques, can operate in a non-linear way. Based on the notion of "structural risk minimization", this algorithm generates vectors which contain the smallest acceptable error limit (Joachims, T., 1998, Text categorization with support vector machines: Learning with many relevant features. In European conference on machine learning (pp. 137-142). Springer, Berlin, Heidelberg.).

Before explaining how the algorithm works, we need to clarify a few terms. One of the basic principles of SVM is called a hyperplane. A hyperplane forms a kind of line inside the object space. This hyperplane is also visible in the accompanying field on each side. The size of the margin (margin) is an important factor. In fact, the more the system finds the hyperplane with the largest margin, the better the results. The hyperplane with the largest margin is called the optimal hyperplane. There are also a number of points in space where the hyperplane is found. However, for a system, the points of interest are those that are on the boundary of the hyperplane. These points are called support vectors (Meyer, D., 2017, Support vector machines. FH Technikum Wien,). Once support vectors are found, other points are ignored (Kotsiantis, 2007). However, finding the optimal linear hyperplane in a large dataset is almost impossible.

ISSN 1512-1232

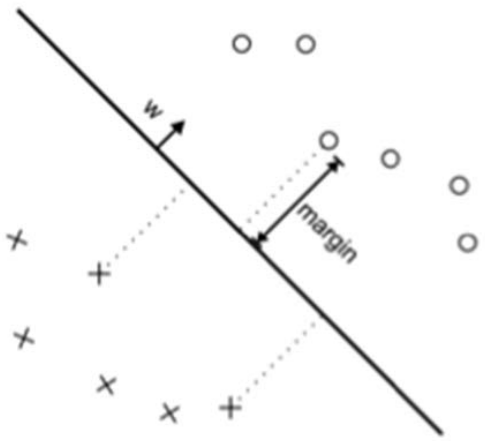

Figure: Representation of the hyperplane and reference (support) vectors

Fortunately, there are two ways to solve this problem. On the one hand, one can add one or more dimensions to the object space. This method ensures that the hyperplane remains linear. On the other hand, if this action is not enough, kernel hints allow to change the shape of the hyperplane. The hyperplane becomes nonlinear. In other words, the shape of the nonlinear hyperplane is determined by the kernel used (Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., &Scholkopf, B., 1998, Support vector machines. IEEE Intelligent Systems and their applications, 13(4), 18-28.). A combination of the two methods is also possible. Despite this, these procedures take time, which means that the SVM algorithm takes much longer to train the data compared to other algorithms.

Unlike most other algorithms, SVMs do not estimate data complexity by the number of features to be taken into account. On the contrary, classification complexity is judged by the size of the hyperplane boundary (Joachims, 1998). Therefore, data containing a large number of features pose few problems for SVMs (Kotsiantis, 2007).

Reality, of course, is much more complex than theory. In practical cases, most data accurately contains a large number of features. This is why SVM can be applied to many real-world applications, such as text classification and processing, facial recognition and fingerprint recognition. Some researchers in the recent past even considered this algorithm to represent the state of the art for that era in the field of pattern recognition (Hearst, 1998; Marrone and Möller, 2011).

In the field of text classification and processing, each document (a tweet in our case) is represented as a vector of words. In addition, vectors usually reflect the frequency of terms present in documents, as well as their distribution across the corpus studied (Hearst, 1998). Language itself has a long list of characteristics. Some even contain more than others. This is why the selection of relevant features is crucial in this area. Choosing the appropriate features leads to a faster and much more efficient system (Hearst, 1998).

ISSN 1512-1232

6. Conclusion

The designed framework collects data from tweets and uses natural language processing techniques to extract features. The natural language processing is then applied to classify the sentiment as positive, negative, and neutral. Polarity and partiality are also calculated by the dictionary, which consists of a semantic evaluation of the tweet. It has been observed that natural language processing is a better method for sentiment analysis than traditional methods. There are some limitations in natural language processing, so other machine learning and data mining techniques may be used in the future to address the limitations of these feature vectors and their selection. Future work will focus on a multilingual machine learning algorithm that processes different types of tasks and easily categorizes data into groups and evaluates them based on real-time data opinions.

Keywords: tweet, sentiment, natural language processing, feature extraction, sentiment classification, machine learning, multilingual machine learning algorithms.

Bibliography

[1] Pang, B., & Lee, L. (2008). *Opinion mining and sentiment analysis*. Foundations and Trends®in Information RetrieJIval, *2*(1–2),1-135.

[2] Крижановский А.А. Автоматизированное построение списков семантически близких слов на основе рейтинга текстов в корпусе с гиперссылками и категориями // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2006». Бекасово, 2006. С. 297-302.

[3] Williams, C. B., & Gulati, G. J. (2009). *Facebook grows up: An empirical assessment of its role inthe 2008 congressional elections*. In Annual Meeting of the Midwest PoliticalScience

[4] Van Hee, C. (2013). *L'analyse des sentiments appliquée sur des tweets politiques: une étude decorpus.* Masterproef. Gent : s.n.,2013.

[5] Tumasjan, A., Sprenger, T. O., Sandner, P. G., &Welpe, I. M. (2010*). Predicting electionswith twitter: What 140 characters reveal about political sentiment*. Icwsm, 10(1),178-185.

[6] Boullier, D., &Lohard, A. (2012). *Opinion mining et Sentiment analysis*. S.l.: OpenEditionPress.

[7] Dave D., Lawrence A., Pennock, D. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. // Proceedings of International World Wide Web Conference (WWW‘03). 2003.

[8] Hu M., Liu, B. Mining and Summarizing Customer Reviews // Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD‘04). 2004.

[9] Jindal N., Liu B. Mining and Summarizing Customer Reviews // Proceedings of National Conference on Artificial Intelligence (AAAI‘06). 2006.

[10] Joachims T. Making large-scale SVM learning practical // In B. Schölkopf, C. J. C. Burges, & A. J. Smola (Eds.), The MIT Press, 1999.

[11] Morinaga, S., Yamanishi, K., Tateishi, K., Fukushima, T. Mining Product Reputations on the Web // In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002). 2002. P. 341-349.

[12] Liu. B. Sentiment Analysis and Subjectivity // In N. Indurkhya& F. J. Damerau. (Eds.). 2010.

[13] Liu, B. Web Data Mining // Springer. 2007. P. 433.

[14] Liu H. MontyLingua: An end-to-end natural language processor with common sense. 2004.

ISSN 1512-1232

[15] Manning C., Raghavan P., Schütze H. Introduction to Information Retrieval. Cambridge University Press, 2008.
[16] Pang B., Lee L. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales // In Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL). 2005. P. 115–124.
[17] Pang B., Lee L., Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques // In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2002). 2002.
[18] Snyder B., Barzilay R. Multiple Aspect Ranking using the Good Grief Algorithm // Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL). 2007. P. 300–307.
[19] Su F., Markert K. From Words to Senses: a Case Study in Subjectivity Recognition // Proceedings of Coling. Manchester, UK: 2008.
[20] Thelwall M., Buckley K., Paltoglou G., Cai D., Kappas A. Arvid Sentiment strength detection in short informal text // Journal of the American Society for Information Science and Technology. 2010. P. 2544-2558.
[21] Turney, P. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 417-424.
[22] Wogenstein F., Drescher J., Reinel D., Rill S., Scheidt J. Evaluation of an algorithm for aspect-based opinion mining using a lexicon-based approach // Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining. 2013.
[23] Ian Goodfellow, Yoshua Bengio, and Aaron Courville «Deep Learning», The MIT Press, Cambridge, Massachusetts, London, England, 2016, 800 pages:

http://faculty.neu.edu.cn/yury/AAI/Textbook/DeepLearningBook.pdf

[24] François Chollet «Deep Learning with Python», by Manning Publications Co., 2018, 386 pages. Print clone:

http://faculty.neu.edu.cn/yury/AAI/Textbook/Deep%20Learning%20with%20Python.pdf

More links

[25] ВААЛ – система контекст-анализа текста. http://www.vaal.ru/
[26] Google Word2Vec. https://code.google.com/archive/p/word2vec/
[27] Sentiment140 – sentiment analysis platform. http://www.sentiment140.com/
[28] Stanford Demo for predicting sentiment of movies reviews. http://nlp.stanford.edu/sentiment/
[29] Yandex Mystem. https://tech.yandex.ru/mystem/
[30] 30dp – opinion search platform. https://www.30db.com/
[31] W2V models – http://panchenko.me/rsr/

Article received: 2025-03-03