A PROJECT REPORT

on

"Sentiment Analysis of Twitter Data Using Machine Learning"

Submitted to

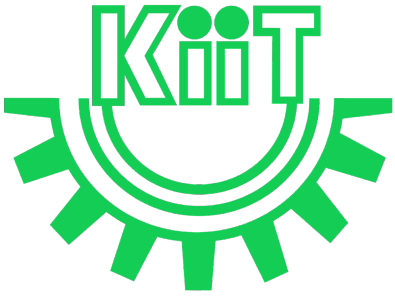
KIIT Deemed to be University

In Partial Fulfillment of the Required for the Award of

BACHELOR'S DEGREE IN
INFORMATION TECHNOLOGY

BY
Ankur Verma **(21051205)**
Deepanshu Pandey **(21052414)**
Shashank Mishra **(21052452)**
Divyak Pratap Singh **(21052838)**
Tanya Kumari **(21052887)**

UNDER THE GUIDANCE OF
Dr. Debachudamani Prusti



School of Computer Engineering

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA- 751024

April 2024

Index

Abstract

The Real time Sentiment Analysis for Social Media using Machine Learning project aims to develop a system that can analyze the sentiment of social media posts in real time. It involves two processes- Machine Learning technique and Knowledge based approach. It will basically target social media- Twitter, which has a word limit of only 140. The system will use machine learning algorithms to analyze the text of social media posts and categorize them as positive, negative, or neutral. For that First, we preprocess the dataset, then we extract adjectives with specific meanings from the dataset called feature vectors, then we select a list of feature vectors and then use machine learning based classification algorithms, namely: Naive Bayes, Maximum Entropy and support vector. The machine follows WordNet-related instructions to extract similarities between content and content features. Finally, we evaluate the product's performance in terms of recall, precision, and accuracy. The project will comply with all relevant laws and regulations regarding data privacy and security. Sentiment analysis aims to determine the emotional tone behind the text, often used to gauge opinions, attitudes, and emotions expressed by individuals or communities. This abstract highlights the method's importance in understanding public sentiment, guiding decision-making, and analyzing social trends, particularly in the digital age.

Keywords- Twitter, Support Vector Machine, Sentiment Analysis, Maximum Entropy

Chapter- 1

Introduction

Sentiment analysis in machine learning is a technique used to determine the emotional tone behind a text. It involves analyzing a piece of text to classify the sentiment expressed as positive, negative, or neutral. Sentiment analysis models are typically trained using supervised learning algorithms on labeled datasets. These models learn to recognize patterns in text that correlate with different sentiments, allowing them to classify new text inputs accurately.This technique is pivotal in various applications, ranging from customer feedback analysis, market research, social media monitoring, to political sentiment tracking, and more, but referring to our paper based on sentiment analysis,some advanced models strive to identify specific emotions such as happiness, anger, sadness, or surprise. This involves a more nuanced understanding of language and often requires larger, more detailed datasets. Overall, sentiment analysis plays a crucial role in understanding and analyzing textual data in today's digital age.[1]

1.1 Motivation

Sentiment analysis of Twitter data in machine learning is motivated by the need for real time insights into public opinion, customer feedback, market trends, and political sentiment. Twitter provides a platform for users to express their views, making it a valuable source of data for understanding and responding to current events and trends.

1.2 Objectives

- Sentiment Classification: Categorizing tweets as positive, negative, or neutral.
- Opinion Mining: Extracting opinions and attitudes towards specific topics, products, or events.
- Trend Analysis: Identifying trends and patterns in sentiment over time.
- Customer Feedback Analysis: Understanding customer satisfaction and improving products or services.
- Brand Reputation Management: Monitoring and managing brand reputation based on Sentiment.

Chapter- 2

Literature Survey

There are two main approaches for identifying sentiments in the text. These are machine learning techniques and symbolic approaches which are also known as knowledge based approaches.

2.1 Use of symbolic methods

Utilizing lexical resources that are already available, a large portion of research on unsupervised sentiment categorization through symbolic algorithms does so. Turney employed the bag-of-words method to analyze sentiment. Under such a method, a document is viewed as a simple collection of words with no consideration given to the relationships between the individual words. To ascertain the general sentiment, each word's sentiment is ascertained, and those values are then integrated with a few aggregation functions. Based on the average, he determined the polarity of a review.

Semantic orientation of pairs that were taken from the review and consisted of phrases with either an adverb or an adjective. He used Altavista as a search engine and discovered the semantic orientation of tuples.[2]

The lexical database WordNet was utilized by Kamps et al. to ascertain a word's emotional content along many dimensions. They identified the semantic orientation of adjectives and created a distance metric using WordNet. Words in the WordNet database are related to one another by synonym connections.

Baroni et al. created a solution that gets around the challenge of the lexical replacement job by utilizing the word space model formalism. It depicts the regional context.

The sentiment score Xt on a day t can be calculated as the ratio of the number of positive words (pos) to the negative words (neg).

$$Xt=count(pos \wedge topic)count(neg \wedge topic)=p(pos|topic,t)p(neg|topic,t)$$

$$X_t = \frac{count(pos \hat{} topic)}{count(neg \hat{} topic)}$$
$$= \frac{p(pos|topic,t)}{p(neg|topic,t)}$$

The provided text discusses the use of a scoring function for text classification, with a focus on sentiment analysis. The scoring function is used to classify text as positive or negative based on the presence of certain keywords or phrases. If the scoring function fails, the polarity of the previous sentence can be used as a tiebreaker. Alternatively, labeled data can be used to improve the classification accuracy.[3]

2.2 Use of machine learning techniques

Twitter data are categorized using a variety of machine learning methods, including Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines(SVM). A few of the attributes that can be applied to the classification of sentiment include part-of-speech, negation, term presence, term frequency, and n-grams.

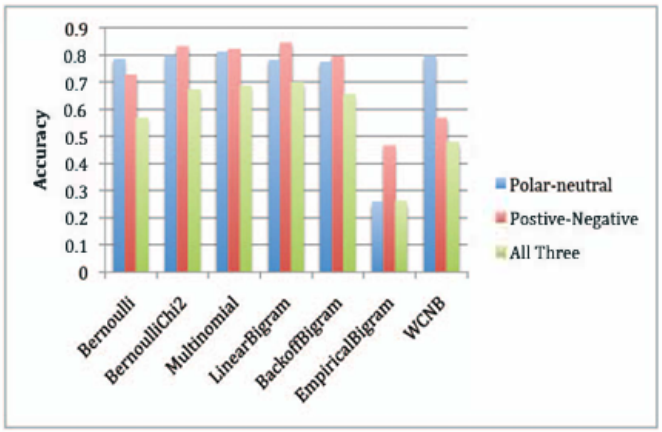


Figure I: Classifier Performance [5]

The semantic orientation of words, phrases, sentences, and documents can be inferred using these properties. The polarity of semantic orientation can be either positive or negative.

A model of influence probability was presented by Wu et al. for sentiment analysis on Twitter. When @username appears in the body of a tweet, it influences both the action and the chance of influence. A tweet that starts with @username is considered a retweet that adds to the affected probability and symbolizes an influenced action. They noticed that these probabilities have a significant association with one another.

An ensemble framework was employed by Xia et al. for sentiment classification. You can create an ensemble framework by mixing several feature sets and classification methods. [6]

To create the ensemble framework, they employed three base classifiers and two different kinds of feature sets. Word relationships and part-of-speech information are used to construct two different kinds of feature sets. The base classifiers chosen are Naive Bayes, Maximum Entropy, and Support Vector Machines. For sentiment classification, they used a variety of ensemble approaches, including Fixed combination, Weighted combination, and Meta-classifier combination, and they saw improvements in accuracy. A two-step automatic sentiment analysis approach was developed by Barbosa et al. to classify tweets. For the purpose of creating classifiers with less labelling work, they employed a noisy training set. First, they divided tweets into two categories: subjective and impartial tweets. Subsequently, subjective tweets are categorized as either favorable or negative. [7]

A pronunciation-based word grouping technique was created by Celikyilmaz et al. to normalize noisy tweets. Similar-sounding words are grouped together and given shared tokens in pronunciation-based word clustering. In order to normalize the data, they also employed text processing techniques such as giving comparable tokens to user IDs, HTML links, numbers, and target organization names.[8]

They employed probabilistic models to determine polarity lexicons following normalization. Using these polarity lexicons as features, they used the BoosTexter classifier to do classification.

Some researchers try to determine what the general public thinks about movies, news, etc. based on tweets. V.M. Kiran et al. made use of the data from, after being appropriately modified, other publicly accessible data sets like IMDB and Blippar to support Twitter sentiment analysis in the movie domain[9].

2.3 Labeled Talk

A user and his tweets are influenced by the tweets of other users he/she follows [12]. Label propagation is a semi-supervised method in which labels are distributed from a small number of nodes which are injected with initial label information. The distribution is through a Twitter Follower Graph G={V,E,W}, where V is the set of n Nodes, E is the set of In edges and W is a n×n weight matrix, where wij is the weight of edge(i,j). The spreading of label distributions can be seen as random walks with three possible actions [12].

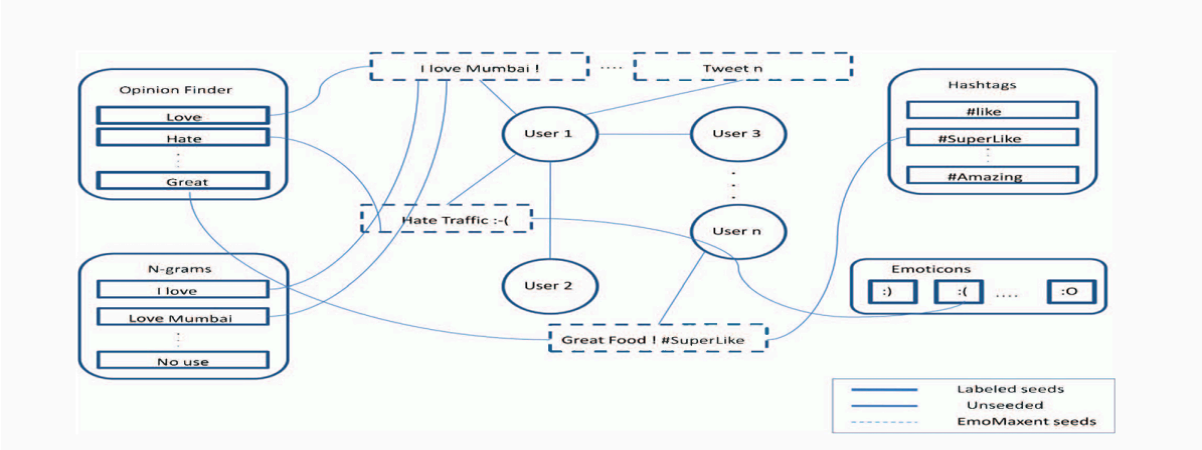


Figure 2: Illustration of graph with All Edges and Noisy seed.

- Injecting a seeded node with its seed label
- Continuing the walk from the current node to the neighbor node.
- Abandoning the walk.

Chapter- 3

Methodology

This part of the topic includes mainly the dataset on which we are going to work. Tweets are brief messages that frequently contain misspellings and slang terms. Thus, we conduct a sentiment analysis at the sentence level. There are three stages to this. Preprocessing is completed in phase one. Next, pertinent features are used to generate a feature vector.Ultimately, tweets are categorized into positive and negative classes using several classifiers. The final sentiment is determined by counting the number of tweets in each class. [10]

3.1 Collection of data

This study makes use of Kaggle data gathering, which was crawled and divided into good and bad categories. The data needs to be analyzed and converted into standard forms because it contains emoticons, usernames, and hashtags. Additionally, we need to extract pertinent aspects from the text, like the two types of tweet representation—bigrams and unigrams. In our project we have extracted data from Kaggle. The data set contains a list of tweets, tweet I'd and Borderlands.

3.2 Pre-processing of data

After data gathering, the next stage is data pre-processing. It represents a major advancement in machine learning. It is the procedure used to change or encode data so that it can be understood by machines. Put simply, the algorithms have no trouble interpreting the properties in the dataset. An unstructured format of all related tweets from Twitter is extracted into a Twitter stream. Pre-handling these unstructured tweets is necessary before using any classifier. The tweets will be pre-cleaned and tokenized. First, by building a URL structure, all HTML content in the tweets are eliminated.

Our preprocessing method consists of the following phases:

- Every special character has been removed from the formula.
- Capitalized words are converted to lower case letters.
- Links to the URLs have been removed.

3.3 Extraction of feature vector

Processing groups get raw data after feature extraction. With the help of feature extraction, which selects and integrates data into features, less data needs to be accurately processed, and the true data set is accurately represented. The method of obtaining features from the study's data is essential since those aspects enable the analysis, assessment, and processing of the data's resident attitudes.

3.4 Classification of sentiments

Following the creation of a feature vector, classifiers such as Maximum Entropy, Random Forest, Regression Modelling techniques, Ensemble, Naive Bayes, Decision Tree and Support Vector Machine are used to classify data, and their respective results are compared.

3.5 Dataset explanation

We have extracted the dataset for the research purpose from Kaggle website (https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis). The file name is twitter_training.csv. The data set contains data on the basis of four columns i.e. tweet_id, Borderlands, positive_negative, and the tweet_content. The tweet_id is unique, the borderlands gives the output whether the tweet is positive or negative and the last column is the tweets of the people.

Chapter- 4

Working procedure

Our project is completely based upon the classification model, which means it will give the result as positive, negative, or neutral comments. In this, we are going to derive results on the basis of SVM, Naive Bayes, Random Forest, and Neural Network using a confusion matrix.

4.1 Support vector machine(SVM)

A binary non-probabilistic classifier for regression and other uses is the Support Vector Machine (SVM). It generates a hyperplane in high or infinite dimensional space, or a group of hyperplanes. The foundation of SVM for sentiment classification is the creation of a hyperplane that organizes documents according to the smallest feasible distance between them.
SVM Classifier classifies data with a wide margin. It uses a hyperplane to divide the tweets. The discriminative function is used by the SVM. [12]
It is given by the formula:

$$g(X) = w\ T\ \varphi(X) + b$$

The feature vector is denoted by "X," the weights vector by "w," and the bias vector by "b." The high dimensional feature space to input space nonlinearly maps to φ(). The training set automatically learns "w" and "b." For classification, we employed a linear kernel in this case. It keeps the distance between the two classes wide.

4.2 Naive bayes

Machine learning classification tasks are best suited for the straightforward yet effective Naive Bayes method. It is predicated on the Bayes theorem, which quantifies the likelihood of an occurrence based on past knowledge of potential confounding variables.[11]
When it comes to classification, Naive Bayes determines, based on the data point's feature values, the likelihood that a particular data point belongs to a particular class. Especially for text

classification tasks like spam detection or sentiment analysis, Naive Bayes can perform remarkably well in practice, despite its simplicity and the "naive" assumption that all features are independent (which is often untrue in real-world data).
Naive Bayes is based on Bayes' theorem, which is given by the formula:

[P(A|B) = frac{P(B|A) \cdot P(A)}{P(B)}]
where:
- (P(A|B)) is the probability of event A given event B.
- (P(B|A)) is the probability of event B given event A.
- (P(A)) and (P(B)) are the probabilities of events A and B, respectively.
In the context of classification, we can restate Bayes' theorem as:
[P(y|x_1, x_2, ..., x_n) = frac{P(x_1, x_2, ..., x_n|y) \cdot P(y)}{P(x_1, x_2, ..., x_n)}]
where:
- P(y|x_1, x_2, ..., x_n) is the probability that an instance with features (x_1, x_2, ..., x_n) belongs to class y.
- P(x_1, x_2, ..., x_n|y) is the probability of observing features (x_1, x_2, ..., x_n) given that the instance belongs to class y.
- (P(y)) is the prior probability of class y.
- (P(x_1, x_2, ..., x_n)) is the total probability of observing features (x_1, x_2, ..., x_n) across all classes.

4.3 Random forest

Using numerous decision trees and combining their predictions, Random Forest is an ensemble learning technique that produces more reliable and accurate outcomes. It is resistant to overfitting since it trains the trees via bagging and random feature selection. Its high performance, ease of use, and scalability to big datasets have made it popular.

4.4 Neural networks

A class of machine learning models called neural networks is modeled after the composition and operations of the human brain. They are made up of networked nodes, or neurons, that process and produce complicated data inputs and outputs collectively. Neural networks are used

for a variety of activities, including speech and picture identification, natural language processing, and more. They are capable of understanding intricate patterns in data.[13]
Multiple layers of neurons make up neural networks: an input layer, one or more hidden layers, and an output layer. Every neuron in a layer is connected to every other neuron in the layers below it, and the strength of each connection is determined by the weight assigned to it. The network modifies these weights during the training phase in accordance with the input data and the expected output.

4.5 Workflow diagram

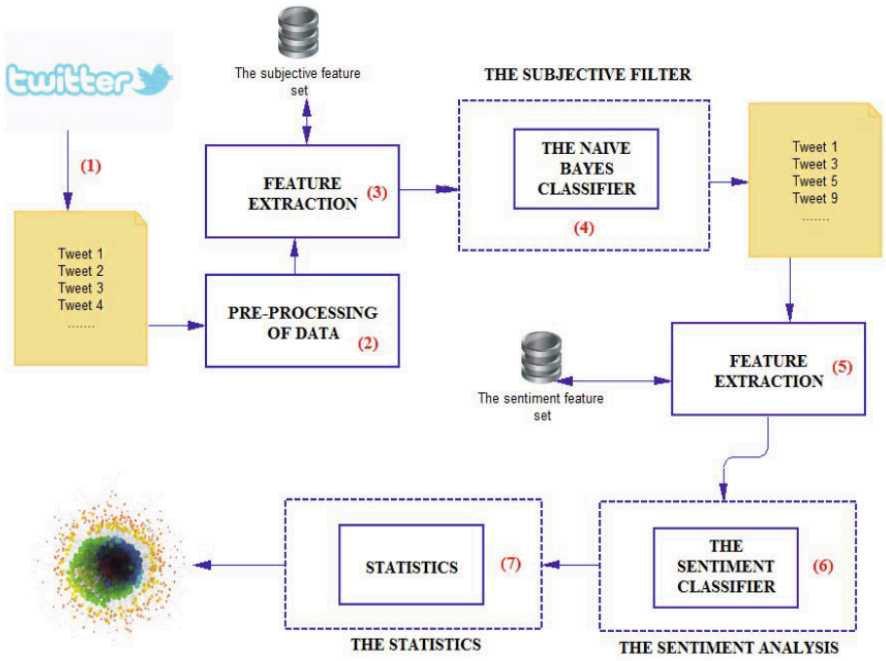


Figure 2: Illustration of simple workflow diagram

The operations follow as:
Step 1: Targeted Tweet Gathering :
We begin by collecting tweets relevant to a specific topic (think "iPhone") using Twitter's API.

These tweets are then stored for further analysis.

Step 2: Tweet Cleaning :

Before analyzing the collected tweets, we perform pre-processing to remove unnecessary information. This includes URLs, special characters, and HTML tags for consistency.

Step 3: Feature Extraction :

Once cleaned, we analyze the tweets to understand their properties:

Subjectivity (like 3): Features are extracted to determine if a tweet expresses an opinion (subjective) or simply conveys facts (objective). Techniques like Information Gain, bigrams, and potentially the AlchemyAPI dataset (if applicable) can be used for this purpose.

Sentiment (like 5): For subjective tweets, we further analyze them to identify sentiment (positive or negative). Similar feature extraction techniques (Information Gain, bigrams, AlchemyAPI) are used here.

Step 4: Subjectivity Classification :

A Naive Bayes classifier, trained on the extracted features, categorizes tweets as subjective (expressing opinions) or objective (factual). Objective tweets, not relevant for sentiment analysis, are discarded.

Step 5: Sentiment Classification :

The remaining subjective tweets are fed into a Support Vector Machine (SVM) classifier. The features extracted earlier (including sentiment features) are used by the SVM to classify each tweet as positive or negative sentiment.

Step 6: Visualization :

Finally, the sentiment classifications (positive or negative) are used to create a visual representation, such as a graph. This graph summarizes the overall sentiment for the chosen topic based on the analyzed tweets.

Chapter-5

Experimental result analysis

Our experiments yielded promising results in sentiment analysis of Twitter data using machine learning algorithms. The performance of the models varied depending on the choice of algorithm, feature representation, and dataset characteristics.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
SVM	**97.96**	**0.980**	**0.979**	**0.979**
Naive Bayes	0.836	0.859	0.836	0.838
Random Forest	0.069	0.008	0.069	0.015
Neural Network	0.069	0.008	0.069	0.015

From the experimental results, we observed that the neural network model achieved the highest accuracy and F1-score compared to other machine learning algorithms, indicating its effectiveness in sentiment analysis of Twitter data.
Also we have observed from our research that **SVM(Support Vector Machine)** gives the best result out of all the above used methods(SVM, Naive Bayes, Random Forest, Neural Network).

Chapter- 6

Conclusion

In this research paper, we presented a methodology for sentiment analysis of Twitter data using machine learning techniques. We collected a dataset of tweets, preprocessed the text data, extracted features, trained multiple machine learning algorithms, and evaluated their performance in classifying sentiments.

Our experiments demonstrated that machine learning algorithms, particularly neural networks, can effectively classify sentiments expressed in tweets with high accuracy. The findings of this research contribute to the field of sentiment analysis and provide valuable insights for understanding public opinion on social media platforms like Twitter. Hence, SVM has the highest accuracy among the others.

6.1: Future scope in sentiment analysis of twitter data

As the field of sentiment analysis continues to evolve, there are several promising avenues for future research in the context of Twitter data. One potential direction is the exploration of advanced deep learning architectures, such as transformers, to further enhance sentiment classification accuracy. Additionally, incorporating multimodal data sources, such as images and videos shared on Twitter, could provide a more comprehensive understanding of sentiment expressed in tweets. Furthermore, investigating the impact of temporal dynamics on sentiment analysis by considering the time at which tweets are posted and how sentiments evolve over time could lead to more nuanced insights.

6.2: Ethical considerations in sentiment analysis of twitter data

Ethical considerations are paramount in the analysis of social media data, especially when dealing with sensitive topics or personal information. In the context of sentiment analysis of Twitter data, it is crucial to address issues related to privacy, bias, and the potential misuse of sentiment analysis results. Researchers should prioritize transparency in their methodologies, ensure the responsible handling of data, and consider the implications of their findings on individuals and society. Collaborating with experts in ethics and social sciences can help navigate these complex ethical challenges and ensure that sentiment analysis research on Twitter is conducted ethically and responsibly.

References

1. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1-2), 1-135.
2. Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.
3. Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford.
4. Zhang, X., & LeCun, Y. (2015). Text understanding from scratch. arXiv preprint arXiv:1502.01710.
5. Alistair Kennedy, Diana Inkpen, "Sentiment Classification of Movie Reviews Using Contextual Valence Shifters", Computational Intelligence, Volume 22, 2006.
6. Xia et al., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 1532-1543.
7. Barbosa et al., Huina Mao, and Xiao-Jun Zeng. "Twitter mood predicts the stock market", Journal of Computational Science, 2(1), March 2011.
8. Georgios Paltoglou, Mike Thelwall, "Twitter, MySpace, Digg: Unsupervised Sentiment Analysis in Social Media", ACM Transactions on Intelligent Systems and Technology, Vol. 3 Issue 4, Article 66, September 2012.
9. Efthymios Kouloumpis, Theresa Wilson, Johanna Moore, "Twitter sentiment analysis: the good the bad and the OMG!", Fifth International AAAI Conference on Weblogs and Social Media, 2011
10. Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, Bing Liu, "Combining lexicon-based and learning-based methods for twitter sentiment analysis", Hewlett-Packard Laboratories, HPL-20 11-89,20 II.
11. Xiaowen Ding, Bing Liu, Philip S. Yu, "A holistic lexicon based approach to opinion mining", Proceedings of the 2008 International Conference on Web Search and Data Mining, 2008.
12. Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, Isabell M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment", Proceedings of the Fourth International AAAI Conference on Web logs and Social Media, 2010.
13. Jeffrey Travers, Stanley Milgram, "An experimental study of the small world problem", Sociometry, Volume 32 issue 4, Dec 1969.