

Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Measurement



Diploma Thesis

**Test Sentence Design
with Minimum Emotional Uncertainty**

Jakub Halcin

Supervisor: Assoc. Prof. Ing. Jan Holub, Ph.D.

Study Programme: Cybernetics and Robotics

Field of Study: Sensors and Instrumentation

May 11, 2014

The thing about perfection is that it's unknowable. It's impossible, but it's also right in front of us, all the time. Kevin Flynn[17]

Acknowledgement

I would like to thank Jan Holub, my supervisor, for his suggestions and constant support during writing of this master thesis. Especially, I would like to thank him for the confidence he gave me during the election decision about the progress of the project.

Prohlášení

Prohlašuji, že jsem práci vypracoval samostatně a použil jsem pouze podklady uvedené v příloženém seznamu.

Nemám závažný důvod proti užití tohoto školního díla ve smyslu §60 Zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon).

In Prague on May 10, 2014

.....



ZADÁNÍ DIPLOMOVÉ PRÁCE

Student: **Bc. Jakub Halcin**

Studijní program: **Kybernetika a robotika**
Obor: **Senzory a přístrojová technika**

Název tématu česky: **Návrh antiemocionálního filtru pro účely subjektivního testování kvality přenosu hlasu**

Název tématu anglicky: **Test Sentence Design with Minimum Emotional Uncertainty**

Pokyny pro vypracování:

Proveďte rešerši dostupných informačních zdrojů v oblasti automatické detekce emocí v textu. Identifikujte metody potenciálně vhodné pro ověření emočně neutrálních vět, vhodných pro subjektivní měření kvality přenosu hlasu např. dle ITU-T P.800. Vybranou metodu aplikujte na dostupná data ze subjektivních testů.


Seznam odborné literatury:

- [1] Tromp, Eric: Multilingual Sentiment Analysis on Social Media. Department of Mathematics and Computer Science, Eindhoven University of Technology, master thesis 2011
- [2] Fink, G.: Encyclopedia of Stress. London, New York: Academic Press, 2007
- [3] McGilloway, Cowie, S., Cowie, R.: Approaching automatic recognition of emotion from voice: a rough benchmark. In: *Proc. of the ISCA workshop on Speech and Emotion*, Ed. Gielen, S. Westerdijk, M. and Stroeve, S., Newcastle, Northern Ireland, 2000, p. 207-212.


Vedoucí diplomové práce: doc. Ing. Jan Holub, Ph.D.

Datum zadání diplomové práce: 23. října 2013

Platnost zadání do¹: 31. srpna 2015


Prof. Ing. Vladimír Haasz, CSc.
vedoucí katedry




Prof. Ing. Pavel Ripka, CSc.
děkan

V Praze dne 23.10.2013

¹ Platnost zadání je omezena na dobu tří následujících semestrů.



MASTER PROJECT ASSIGNMENT

Student: **Bc. Jakub Halcin**

Study programme: **Cybernetics and Robotics**
Specialisation: **Sensors and Instrumentation**

Title of Master Project: **Test Sentence Design with Minimum Emotional Uncertainty**

Návrh antiemocionálního filtru pro účely subjektivního testování kvality přenosu hlasu
(in Czech)

Guidelines:

Make yourself acquaint with automatic text emotion detection methods. Identify possibly suitable methods to validate emotionally neutral sentences used for subjective testing of transmitted speech quality, e.g. according to ITU-T P.800. Apply the selected method to data available from subjective tests.

Bibliography/Sources:

- [1] Tromp, Eric: Multilingual Sentiment Analysis on Social Media. Department of Mathematics and Computer Science, Eindhoven University of Technology, master thesis 2011
- [2] Fink, G.: Encyclopedia of Stress. London, New York: Academic Press, 2007
- [3] McGilloway, Cowie, S., Cowie, R.: Approaching automatic recognition of emotion from voice: a rough benchmark. In: *Proc. of the ISCA workshop on Speech and Emotion*, Ed. Gielen, S. Westerdijk, M. and Stroeve, S., Newcastle, Northern Ireland, 2000, p. 207-212.

Master Project Supervisor: **Assoc. Prof. Ing. Jan Holub, Ph.D.**

Valid until: **August 31, 2015**

L.S.

Prof. Ing. Vladimír Haasz, CSc.
Head of Department

Prof. Ing. Pavel Ripka, CSc.
Dean

Prague, October 23, 2013

Abstract

This diploma thesis analyzes the influence of sentiment contained in the text to the results of transmitted voice quality subjective testing as per ITU-T recommendation P.800. As part of the thesis, process to eliminate presence of sentiment in source sentences is designed. The creation of multilingual corpora and implementation of classifiers based on the principle of maximum entropy is described in detail. Results of the work and their statistical evaluation are presented in the context of real data from a corporate test file.

Keywords: sentiment analysis, machine learning, subjective voice quality testing, data mining

Abstrakt

Tato diplomová práce se zabývá analýzou vlivu sentimentu obsaženého v textu na výsledky subjektivního testování přenosu kvality hlasu podle doporučení ITU-T P.800. V rámci práce je navržen postup eliminující přítomnost sentimentu ve zdrojových větách používaných jako zdrojové pro toto testování. Detailně je popsána implementace klasifikátoru na principu maximální entropie a tvorba vícejazyčných podpůrných korpusů. Výsledky práce jsou zasazeny do kontextu s daty z reálného průmyslového testovacího souboru a jsou statisticky zpracovány.

Klíčová slova: sentimentální analýza, strojové učení, subjektivní testování kvality hlasu, dolování dat

Contents

1	Introduction	1
2	Problem definition	3
2.1	ITU-T recommendation requirements	3
2.1.1	ITU-T E-model	3
2.1.2	ITU-T expectation factor	4
2.1.3	The current challenge	5
2.2	Emotions in text	5
2.2.1	Emotion and feeling	6
2.2.2	Outward and inward sentiment	6
2.3	Subjective measurement transmission quality	6
2.4	Meaning of opinion score and subjective testing	7
2.5	The main aim and potential risks	7
3	Existing method comparison	9
3.1	Decision tree learning	9
3.2	Naive Bayes classifier	10
3.3	Maximal Entropy classifier	11
3.4	Softmax classifier	11
3.5	Support vector Machine	12
4	MaxEnt Classification	14
4.1	Terms definition	14
4.2	Model description and conditions	15
4.3	Searching for model	16
4.4	Software implementation	17
4.4.1	Sharp entropy	17
4.4.2	Implementation of the training process	17
4.4.3	Implementation of the classification process	17
4.4.4	Software output explanation	18
5	Corpus and dataset	19
5.1	Existing corpora	19
5.1.1	Sentiment140 twitter corpus	19
5.1.2	MLSA German corpus	20
5.1.3	MPQA opinion Corpus	20
5.1.4	The MICRO-WNOP Corpus	20
5.2	Creating new corpora	21
5.2.1	Multilingual corpus needs	21
5.2.2	Data mining	22

6	Metrics used in the project	23
6.1	Confusion matrix	24
6.2	F1 measure classifier sensitivity	24
6.3	ANOVA	25
6.4	Matthews Correlation Coefficient	25
6.5	ROC curves, AUC	25
6.6	Crossvalue validation	27
7	Project results	29
7.1	Training corpora	29
7.1.1	Czech corpus	29
7.1.2	English corpus	29
7.2	Project metrics fulfillment	30
7.2.1	Classifier confusion matrix	30
7.2.2	ROC analyses, AUC	30
7.3	Subjective measurement data and classifier results	31
7.3.1	Correlation	32
7.3.2	ANOVA single Factor	32
7.3.3	Live testing data behavior	33
7.4	Repeatability of results	39
8	Conclusion	44
8.1	Summary	44
8.2	Benefits for future projects	44
8.3	Personal benefits	45
	Bibliography	46

List of Figures

2.1	Draft of complete E-model and it's parameterization	4
6.1	ROC ideal classified	26
6.2	ROC random classified	27
6.3	ROC dysfunctional classified	28
7.1	ROC of implemented classifier	31
7.2	MOS score for neutral and sentimental sentences, very low quality network . .	34
7.3	Variance of MOS score for neutral and sentimental sentences, very low quality network	34
7.4	MOS score for neutral and sentimental sentences, low quality network	35
7.5	Variance of MOS score for neutral and sentimental sentences, low quality network	35
7.6	MOS score for neutral and sentimental sentences, medium quality network . .	37
7.7	Variance of MOS score for neutral and sentimental sentences, medium quality network	37
7.8	MOS score for neutral and sentimental sentences, upper medium quality network	38
7.9	Variance of MOS score for neutral and sentimental sentences, upper medium quality network	38
7.10	Example of MOS cultural dependency (img. source [4])	39
7.11	Subjective test repetition Sample A, PH1 and PH2 MOS	41
7.12	Subjective test repetition Sample A, PH1 and PH2 MOS variance	41
7.13	Subjective test repetition Sample B, PH1 and PH2 MOS	42
7.14	Subjective test repetition Sample B, PH1 and PH2 MOS variance	42

List of Tables

2.1	The relationship between R-factor and MOS	4
7.1	Confusion matrix of sentiment classifier	30
7.2	ANOVA summary for MOS variance and Sentiment score	32
7.3	ANOVA results for MOS variance and Sentiment score	33

Chapter 1

Introduction

This thesis is based on studying research papers, previously developed thesis and testing. Thesis is intended as research thesis. The main objective was to find neutral sentences useable for subjective testing of voice quality. This will lead to increase confidence in the results of subjective tests. My goal was to compare relevant methods of detection sentiment in text and chose one of them, which will be implemented and tested.

Sentiment analysis theme is very hot for marketing and investment use. But this thesis comes with solid development. It proposes a classification method which is able to provide sentiment metrics for improving results of subjective testing of voice quality. It's currently totally new theme. When I tried to search answers for my questions it usually led to big number of completely new questions. Some of them I have overcome by own ideas, many of them is still opened for the future research.

This thesis is divided into eight chapters.

After this introduction, the second chapter deals with a deeper understanding of the subjective tests issues. There is also disclosed identification of the non-technical problem and its transfer to the solution cybernetics and measurement problems. The basic hypothesis and requirements for successful completion of this project are established. The chapter is concluded with a summary of the objectives and potential risks of the project.

In the third chapter, there are existing methods which can be used as a partial solution to the described problem. These methods were qualitatively explored and the best option was chosen in terms of real achievement of the desired objectives.

The method of the solution algorithm is maximal entropy classifier. It is closely described in the fourth chapter. There can be found mathematical derivation, optimization algorithm and description of the software implementation. Maximal entropy algorithm has become part of the application designed for laboratory using.

The following fifth chapter deals with the formation of the corpus. In the introduction, existing solutions seemingly satisfactory classification problem are compared. Requirements classifier constructed on the training set are defined and all examined corpora are proved as inadequate. The chapter continues the idea of creating a new corpus based on data mining

public sources.

The sixth chapter deals with the introduction of metrics by which success rate of project should be assessed. There are commonly used indicators and tools that can provide a true imagination of the effectiveness of binary classifiers.

The most important seventh chapter is devoted to the fulfillment of project metrics. Specifications multilingual corpora, which I created, can be found here. There is classifier functional verification of statistical point of view and it is provided with a direct view of the results.

There is separate place in the seventh chapter for the observations associated with corporate data from Audience Inc. It is the original developer hypothesis - sentiment contained in the test sentences used for subjective testing of quality voice transfer testing increases the quality score. However, this could not be confirmed and hypothesis is not valid for obtained data.

Chapter 2

Problem definition

2.1 ITU-T recommendation requirements

Initialization impulse to the preparation of this thesis was to find a method which requests improving for ITU-T recommendation P.800 [34]. This recommendation relates to the methodology of the subjective voice quality testing various communication channels. Here are a few terms that you need to know for understand this thesis issue.

2.1.1 ITU-T E-model

ITU-T G.107 describes more closely model paths for voice transfers. It is well known as the E-model [5]. Main output of E-model is the scalar rating factor, which can be seen as metric of subjective rating of voice transfer quality.

$$R = R_0 - I_s - I_d - I_{e-eff} + A \quad (2.1)$$

Rating factor (equation 2.1) is calculated as the sum of R_0 signal-to-noise ratio, compensation constant A , which is important as impairment factor (so called Advantage factor or Expectation factor). From these values is subtracted depreciation factor due to delayed transmission I_D , I_{e-eff} depreciation factor by codec settings (Equipment impairment factor), and all other defects I_s (Simultaneous impairment factor). Complete E-model principle can be seen in the figure 2.1, complementary parameters includes.

In practical case E-model can be used for estimation of voice transfer quality in developed transfer networks. Recommended table for quality classification exist under the E-model. It provides values Rating factor.

How to use E-model is presented in the figure 2.1. Signal-to-noise ratio and compensation constant, depreciation factor and other defect constant can be computed, others not. Actually, A-factor and I_e effect of equipments is necessary to find in reference tables.

The second typical usage is measuring parameters existed and working (live) networks. In this case parameters are derived from network features and reference tables. As stated in the materials used in the practice, the values of parameters are very approximate and they

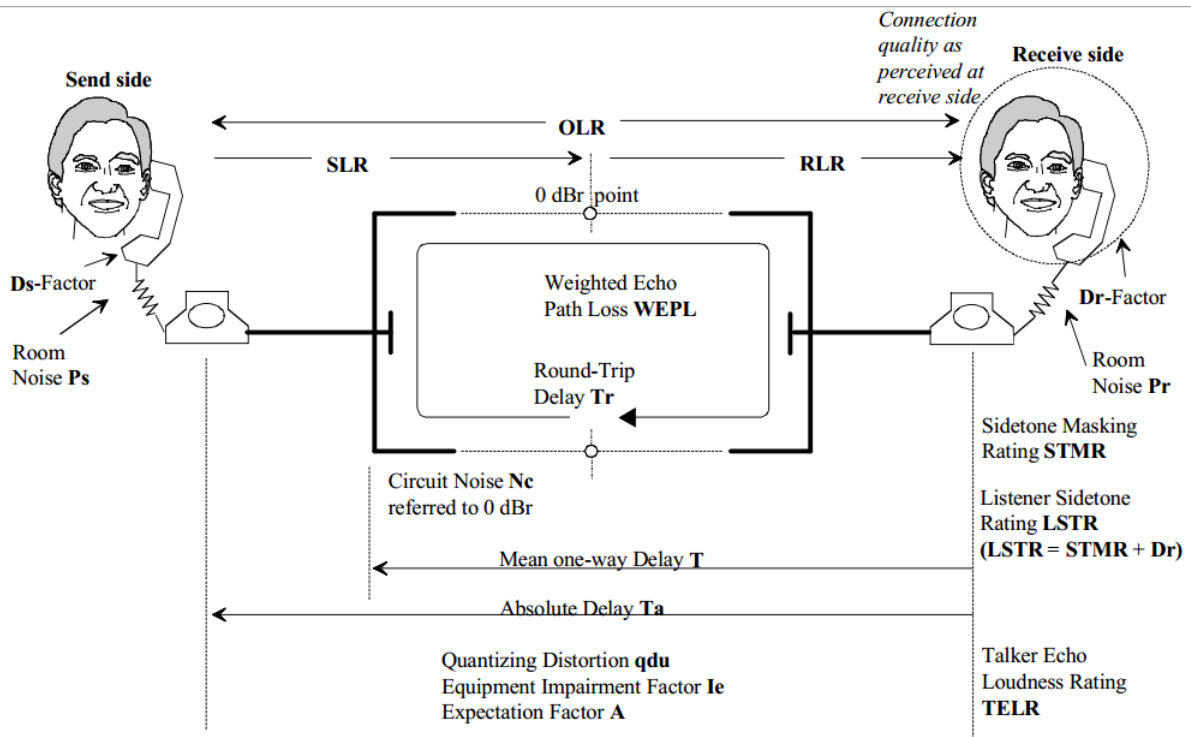


Figure 2.1: Draft of complete E-model and its parameterization

User Satisfaction	Rating factor value range	MOS equivalent
Very satisfied	90-100	4.3 - 5.0
Satisfied	80-89	4.0 - 4.3
Some Users Dissatisfied	70-79	3.6 - 4.0
Many Users Dissatisfied	60-69	3.1 - 3.6
Nearly all Users Dissatisfied	50-59	2.6 - 3.1
Not recommended	less than 50	1.0 - 2.6

Table 2.1: The relationship between R-factor and MOS

don't often correspond to the actual parameters of the network. The most important thing is the measurement and evaluation Mean opinion score (MOS 2.4) parameter. MOS is the output parameter of subjective testing existing physical network. Its expected value can be approximately derived from R-factor according to the table 2.1.

2.1.2 ITU-T expectation factor

This section is focused on the A-factor. It is interesting; the authors of this recommendation indicate that A factor is not important in the E-model if the E-model (2.1.1) is used in independent purpose. This information is based on 15 years experiences authors of this recommendation. Notice A Factor enhances Rating factor and it isn't negative number (usually used values chap. 3.6 at [33]).

Let's summarize a presumptions that the A factor is a figment of human emotions. When

working with E-model, it is necessary to deal with their origin. Emotions can affect the testing and significantly increase the rating factor. It's not clear how much is A-factor dependent on sentiment and what kind of sentiment is deciding.

The current developers working with the E-model to explain expectation value factor as improving enthusiasm. It can be presented as a fact person performing subjective testing has previous experience of older and less quality technology. So, tester may be affected by enthusiasm for new technology with better quality and MOS (section 2.4) is artificially increased. But there are probably several other components of A-factor which haven't been sufficiently discussed yet.

2.1.3 The current challenge

In 2013, it was claimed that the A factor is dependent on sentiment and it was come through sentimental test sentences into MOS. This assumption was spoken, but it wasn't sufficiently tested and currently can be a source of discrepancies in the evaluation of the results of subjective testing.

My approach should provide a method that allows to deciding sentiment hidden in the test sentences. It is currently generated on the basis of statistical processing of phonetic form of the language. According to developed methods it will be investigated dependence of the results of subjective tests and the degree of sentiment in sentences that were used for these tests.

For future development is considered elimination sentiment in two steps. The first is generating a significantly higher number of sentences. First step will be filtering-out sentimental sentences and on the out of filter we will have several sentences with very low probability sentiment rate. The second step will be assess the phonetic interpretation of a sentence sentiment content by checking audio post processing methods.

The theme of this thesis is closely interlinked with the first step - automatic text classification by sentiment to obtain neutral sentences and how/whether it improves reliability subjective tests.

2.2 Emotions in text

There are several abstract phrases with very similar meaning. In the English terminology, we can talk about feelings, emotions and sentiment. An explanation of these concepts certainly does not fit into a standard metrology. For understanding, we need deal with psychology and philosophy.

To consolidate the terminology I summarize the basic definitions of these terms by M/C Journal [29]. These terms will be used in the same significance in this thesis.

2.2.1 Emotion and feeling

An emotion is the projection/display of a feeling. Unlike feelings, the display of emotion can be either genuine or feigned.

A feeling is a sensation that has been checked against previous experiences and labeled. It is personal and biographical because every person has a distinct set of previous sensations which proceed from when interpreting and labeling their feelings.

2.2.2 Outward and inward sentiment

A sentiment is impression which can be composed of two parts. it fits to dividing according to D. Hume analysis [16]. In this thesis we can find outward and inward sentiment.

Paraphrase D. Hume (philosopher):

Outward sentiment is an impression of something from the external world, obtained through one of the five senses.

Inward sentiment is an impression which is of something not from the external world; for example nervousness or anger.

So our text sentimental analysis is actually examining the outward sentiment. The results subjective testing will consist of two components. We need to realize it is related to the content of the text, but also the pre-stimulation of the person who assesses the text is important.

Quoting Doc. RNDr. Josef Jelen, CSc. (physicist):

Sentimental is everything, what seems to be a person of interest and it evokes emotion.

The forming sentences can be regarded as a random process. Initially we have empty set, which is neutral. By gradually adding words we can change the probability that the context will be neutral or sentimental. It is noteworthy that adding the word, sentence can be easily changed from neutral to sentimental. But reverse procedure does not work.

2.3 Subjective measurement transmission quality

Upper described E-model (section 2.1.1) is put to context of methodic recommendation for subjective and objective testing quality transmission in telephony. For better understanding I focused on the E-model environment and its details. It's fully described in recommendation ITU P.800.

2.4 Meaning of opinion score and subjective testing

If we speak about subjective testing we mean methods for testing quality of voice transfer networks. That could be applied for testing and comparison technology in telephony for ex. satellite, GSM or VOIP transfer sites. Parallel to subjective testing there are also objective testing technologies, which aren't part of this thesis.

Note it can be said subjective testing of voice transfer quality requires statistical significant number of professional testers and environment. Subjective testing of voice transfer quality is described in the P.800 recommendation. This is method used for decades and it's updated for modern networks.

Complete subjective testing by P.800 covers generating test sentences and recording samples by profession speaker. Core of test is based on performance prepared samples over voice transfer network. Every sample is evaluated by professional listeners. Listeners choose the values on a predefined scale (from 5 to 1) by individual opinion. All individual scores are computed over arithmetic mean and output is called MOS. It is scalar quantity without dimension.

Because MOS is subjective measured parameter, it can be different for the same network system when it is used by different users. As it can be read in the paper [4] MOS variation is generated by cultural and individual specification. More about this problematic can be read in the section 7.4. Competitive objective test provides better reproducible results but it doesn't say how satisfied users of the network are.

2.5 The main aim and potential risks

Aimed at eliminating sentiment in subjective testing quality of voice transfer, some fundamental questions raise.

If we aim at sentiment function in subjective testing and its elimination, we can expect several problems. In view of the goal, we have no guarantee; sentiment will lead to decrease significantly improving reliability of testing? Before starting work on this thesis, there was only said idea.

Ideally, it can be used automated classification to select the sentences with absolutely no sentiment. If we choose to use automated classification - what is the best classifier for natural language sentiment? And, will we be able to find a suitable training dataset for such a classifier?

We have the ability to use several types of classifiers. However, even if we manage to find the perfect classifier, this classifier will only be as good as the training set. Training set or corpus must be in good relation with the source phrases used for subjective testing.

There are others risks. The words in the training set may be many, but it cannot cover the whole vocabulary of the language. It is therefore possible that information in the training set isn't enough to decision if sentence in input contained by sentiment. Bad case may

be evaluation of sentiment information obtained from extensive context on too few words. Results may be completely misleading.

The appalling remark can also be found in [1]. Author of this paper argues that there is also a kind of emotion that can be considered neutral. If this is true, we will not be able to distinguish neutral sentiment in text and text contained by neutral emotions.

I tried to find answers to these questions and summarize their findings in the following chapters.

Chapter 3

Existing method comparison

In this chapter I was focused on choosing useable type of classifiers for later implementation. Firstly I found there are two basic approaches. Developers of sentiment analysis utilities often choice between lexical decision tree and supervised machine learning. At first sight I studied existing project, their technologies and achievements. I tried to consider possible accuracy, nature of word processing and demand of implementation. Critical factor is practical feasibility in time horizon this thesis.

3.1 Decision tree learning

This classification method is based on decision tree building. It's learning with supervisor. All training data is bootstrapped with replacement. Next step is randomly selected n features without replacement and k threshold values. Thresholds are chosen between minimal and maximal value.

After that, we need to choose value with minimal Gini coefficient. Gini index (known as Gini coefficient in prediction theory, too) is a value, which we can get from Lorentz diversification. How to do that, it can be found in the article [10]. It isn't necessary to create self implementation, because there are ready to use APIs. I have not found API useable in .Net C sharp projects, but there is for example Open MPI API working under JAVA, Python, Matlab or R-studio projects.

Crucial step is creating nodes of decision tree. Exact results text sentiment analysis of this method can be found in [9] and precision of these practical tests is no more than 75 % with real conditions.

Advantages:

- easy for effective implement with parallel computation

Disvantages:

- small classification accuracy
- very limited gain information from the training set [20]

3.2 Naive Bayes classifier

Naive Bayes (very similar to classifier under short MDT) is one of most used classifier. It's very popular because implementation is easy. Naive Bayes is classic two step classifier - constituted from training and classifying algorithm. All theory over this classifier is origin derived from Bayes' theorem (equation 3.1).

$$P(W|L) = \frac{P(L|W)P(W)}{P(L)} \quad (3.1)$$

Training in this case is performed as posterior probability estimation for every collected n-grams. It could be written as equation 3.2.

$$posterior = \frac{prior \prod likelihood}{evidence} \quad (3.2)$$

There is one strong precondition. It is required that the individual n-grams were conditionally independent. It can be intuitively to sense that this assumption is not valid. Words or n-grams interact with each other and their sentiment isn't independent.

The second phase - algorithm for classification is based on likelihood maximalisation. So object is classified to class with maximum posterior likelihood. This result can be obtained from the equation 3.3.

$$class(c1, \dots, c2) = argmax_{c \in C} (P(c|d)) = argmax \left(P(c) \prod_{1 \leq k \leq n_d} P(t_k|c) \right) \quad (3.3)$$

Notation by the article [13], where t_k are the tokens (terms/words) of the document, C is the set of classes that is used in the classification, $P(c|d)$ the conditional probability of class c given document d , $P(c)$ the prior probability of class c and $P(t_k|c)$ the conditional probability of token t_k given class c .

Although the assumption of independence is not met, Naive Bayes classifier still provides acceptable results in the classification of natural text. Typical application of Naive Bayes classifier can be found in the article [28]. After exploring the previous work I have to say, for sentiment analysis Naive Bayes is usable. But it doesn't achieve the best results.

Exact results of text sentiment analysis by Naive Bayes classifier can be found in [23] in table 1. It could achieved precision from 60 % to 80 %, improved adapting Naive Bayes algorithm described by [30] achieved 83 %.

Advantages:

- easy for implementation
- very low computer performance demands

Disvantages:

- precondition isn't clearly valid
- bad mutual information between objects

3.3 Maximal Entropy classifier

Due to the minimum assumptions the Maximum Entropy classifier (well known as MaxEnt) we regularly use it when we don't know anything about the prior distributions and when it is unsafe to make any such assumptions. Moreover Maximum Entropy classifier is used when we can't assume the conditional independence of the features. These are described in the [35].

The name is slightly misleading though. Not used directly classic formula of maximum entropy, but entropy model. Theoretically, it is necessary a priori knowledge of probability. In our case, the probability of occurrence of words in sentences with neutral and emotional sentiment.

In a real situation, we don't know a priori probability. So we can only estimate on frequency of occurrence of words acquired earlier measurements.

Thanks to classification into two classes, we can successfully exploit mutual information between the statistics for individual words. Solution of optimization problem is followed, but it is for filtering into two categories already described and successfully resolved.

The output of this process is a classifier that processes the text specified in its input. At the output, we get the probability if input is neutral or sentimental type of text. A complete theory require to build entropy model, implementation of a classifier, finding suitable training set is described in detail in chapters 4 and 5.

The exact results text sentiment analysis of this method in multilingual sentiment text classification can be found in the [7]. This type of classifier usually achieves over 90 % accuracy and it doesn't suffer from overfitting effect.

Advantages:

- great results in natural language processing
- maximizes the use of information in the training data

Disadvantages:

- need for quality training set
- "blackbox" features, not well described [31]

3.4 Softmax classifier

It is a regression model which generalizes the logistic regression[12] to classification problems where the output can take more than two possible values. Many existing algorithms distinctive kinds of emotions in the text are based on this type of classifier. Using logistic regression assumes the dependent variable with a binomial distribution.

Most implementations could be imagined as neural network where classifier is interpreted as one level neurons. There are used lower level neurons as inputs and output is vector class probability. The classifier needs training with a teacher.

$$J(\theta) = -\frac{1}{m} \left[\sum_{i=1}^m \sum_{j=1}^k 1 \{y^{(i)} = j\} \log \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^k e^{\theta_l^T x^{(i)}}} \right] \quad (3.4)$$

Training phase is based on working with Softmax cost function (equation 3.4 with notation by Stanford UFLDL Tutorial) and optimization to minimal value. It requires logistic regression. Each variable is assigned to only one category. Concerning the smoothing and classification of one variable can be simultaneously assigned to N categories. It is necessary to create a cost function of N and N separated Softmax classifiers.

We should note that Multinomial Logistic Regression is closely related to MaxEnt algorithm because it uses the same activation functions.

The exact results text sentiment analysis of this method can be found in the [3]. Authors described development, testing and results of Softmax classifier for identifying type and strength of emotions. There isn't virtually reason to compare precision to others methods. Because, mostly released projects powered by Softmax classifier are closed to multidimensional sentiment finding and their results are not comparable to two class classification.

Advantages:

- we don't have to worry as much about features being correlated
- after completion of the learning it has low demands on CPU time and RAM
- works well with semi-continuous features or with different types object

Disadvantages:

- requires significantly more time to be trained comparing to Naive Bayes
- optimization needed
- returns bad results for correlated features (multicollinearity problem [36])

3.5 Support vector Machine

Well known by the acronym SVM, it is a machine learning method based on linear algebra. The training data are generated the vectors symptoms. Result of SVM training is to find the optimal furthest hyperplane. How to choose the separating hyperplane was proposed by Vapnik and Lerner[8] and they put base proof for SVM (equation 3.5).

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\frac{h(\ln(2l/h) + 1) - \ln(\eta/4)}{l}} \quad (3.5)$$

Note, there is $R(\alpha)$ as the actual mean error of model, l is data length, α are defined parameters of model and $R_{emp}(\alpha)$ is estimation of the mean error based on empirical data (training dataset). In the next step, we need to establish loss function which leads us to get empirical risk.

There isn't easy way to resolve optimization problem. So, SVM methods rely on dual problem resolution. Moreover, it shows that for practical applications aren't possible to operate with simple hyperplanes because we can't separate classified classes ideally. It goes to algebraic hypersurfaces.

Practical results are quite competitive to MaxEnt classifier. SVM usually achieves an accuracy of around 90% and in the academic environment is very popular. Its advantages can be attributed to the algebraic derivation function. MaxEnt is a blackbox compared to SVM algebraic mathematics.

Advantages:

- optimality problem is convex[15]
- high accuracy
- full algebraic inference

Disadvantages:

- memory-intensive
- harder for implementation than MaxEnt
- non-parametric SVM is blackbox like MaxEnt

By the way, it could be interesting imagination SVM classifier in OLAP cubes database. This method should be able to benefit from big data processing in data centers and it should be pretty easy for implementation in n-dimensional vector cubes. Unfortunately, I haven't found mention that anyone would have tried it so far.

Chapter 4

MaxEnt Classification

In this chapter I have focused on Maximum Entropy classifier theory. It describes the operating principle of classifier usable for later software implementation.

I'm very interested in this type of classifier because it provides very useful results in sentimental analysis. If we limit ourselves for working with a small number of n-grams, this method is friendly to the computational demands. It doesn't put any fundamental requirements for data, so we thanks to the deployment, we will not make a major failure. And above all, MaxEnt is based on a very strong idea that allows us to maximum utilization of small training set.

At start I was very inspired by [6] Adam Berger(1996) Maximal Entropy Tutorial.

4.1 Terms definition

Let me introduce definition of variables used in this chapter.

- x_i - contextual information object, should be imagine as sets of word (input)
- y_i - set of classification classes (neutral and sentimental texts)
- w_k - unit of contextual information object (should be imagine as one word)
- c_i - one of defined classes
- $E(f_i)$ - expectations of features (actual predicted count of feature hits)
- $Z(x)$ - normalization constant (length of training dataset)
- λ_i - parameter of probabilistic model
- $p(a)$ - probability of event a
- $\tilde{p}(a)$ - estimation probability of event a

$f_i(x, y)$ is feature function. It has two inputs (described above) and one two-state output. Returns 1 iff object x contains word w_k and x context is trained as required class c_i . Returns 0 otherwise. Features are elementary pieces of evidence that link aspects of what we observe

with a class that we want predict.

Note most of above designations were taken from the A. Berger texts [6] and it is commonly used in the text of his followers without direct explanation.

4.2 Model description and conditions

The aim of the MaxEnt classifier model is to find the probability distribution that satisfies the principle of maximum entropy. This principle can be interpreted as the fact if the model has little information, then it uses the most probable shape of the probability distribution and exactly according to the information that the algorithm has. So, try to find such a probabilistic model, which has the maximum entropy.

Principle of maximal entropy is often expressed by equation 4.1.

$$\operatorname{argmax} \left(- \sum_{i=1}^n p_i \cdot \log_2 \cdot p_i \right) \quad (4.1)$$

The first step is the estimation of the training set. Gradually, estimation the empirical probability distribution is built. It can be done according to the equation 4.2.

$$\tilde{P}(x, y) = \frac{1}{Z(x)} \operatorname{freq}(x, y) \quad (4.2)$$

The obtained empirical probability distribution of feature function $f_i(x, y)$ is applied. It should be written as:

- $f_i(x, y) = 1$ iff word x is in class y
- $f_i(x, y) = 0$ iff word x isn't in class y

Combination of feature functions $f_i(x, y)$ and estimation from empirical probability $\tilde{P}(x, y)$ we get constrains for our probability model. From maximum entropy models we choose models declared by the equation 4.3.

$$p(y|x) = \frac{1}{Z(x)} \exp \left(\sum_{i=1}^n \lambda_i f_i(x, y) \right) \quad (4.3)$$

It's called log-linear model equation.

The searching for the probabilistic distribution is subject to additional conditions. These conditions are created from expected value feature function (equation 4.4).

$$\sum_{x,y} p(x, y) f_i(x, y) = c_i \quad (4.4)$$

This expected value can be obtained from $\tilde{p}(x)$ relative frequency of context object in training dataset. So, let's denote the frequency of feature c_i . It is a constant. Make estimation by equation 4.5.

$$\sum_{x,y} p(x,y) f_i = \sum_{x,y} \tilde{p}(x) p(y|x) f_i(x,y) \quad (4.5)$$

As noted Vasilis Vryniotis(2013) [35], majority parts of pairs (x,y) should be included only once in the training set. So, this lead to predominance of zeros values for all elements that are not found in the training set and value $\tilde{p}(x,y) = \frac{1}{Z(x)}$ for most others. This is due to the fact that the MaxEnt does not work with any information that is not contained in the training set.

4.3 Searching for model

Now we need denote model parameters (equation 4.8). There are several different algorithms which can found parameters for generation model [19]. One of the simplest algorithm is hidden under acronyms GIS (Generative Iterative Scaling). In the design of the algorithm I primarily benefit from the experience of the authors of article [14]. GIS is based on iteration over model parameters until its convergence

We start with constrains substitution by the equal 4.6.

$$\sum_{x,y} \tilde{p}(x) p(y|x) f_i(x,y) = E f_i \quad (4.6)$$

$$C = \sum_{i=1}^n f_i(x,y)$$

C is the total number of features which are active for a particular (x,y) pair. It may be replaced by a constant which is equal to maximum of C (by equation 4.7).

$$C_{max} = \max_{x,y} \sum_{i=1}^n f_i(x,y) \quad (4.7)$$

It's described in detail in [19]. Select any value for start lambda parameter (i.e. λ). Initial value λ_0 is chosen as zero.

$$\lambda_i^{t+1} = \lambda_i^t + \frac{1}{C} \log \frac{E(f_i)}{E(t)(f_i)} \quad (4.8)$$

Transferred to differential form (equation 4.9).

$$\Delta\lambda = \frac{1}{C} \log \frac{E_p f_i}{E_{p(t)} f_i} \quad (4.9)$$

Then repeat while λ non-convergence calculation next λ steps from equation 4.6.

By introducing a correction function (4.10),

$$f_c(x,u) = C - \sum_{i=1}^n f_i(x,y) \quad (4.10)$$

it can be bypassed the requirement for the constancy of the sum all feature values. Berger showed that the drop can also use the correction function, while GIS is convergent.

4.4 Software implementation

When I decided for using MaxEnt I tried many Maximal Entropy classifier implementations. Several of them are commercial projects, so they aren't useable for our development. However, I've found a way how to use the API with self-expansion and changes.

4.4.1 Sharp entropy

My final implementation is based on the package Sharp entropy. This is a project prepared under .Net C and it is licensed under the LGPL. So it is possible to inherit classes for generating MaxEnt model and use it for further modification. There are two basic inherited classes - Model and Trainer.

4.4.2 Implementation of the training process

The algorithm starts with training. At the entrance is required to have already prepared the training data, which are always classified into two classes - a sentimental or neutral. Before using descendant class myTrainer, it's needed to tokenize data. It is also expected that the punctuation marks or whitespace characters have been removed from training tokens. Additionally, words without sentiment meaning are also removed from used tokens.

Before putting the data to core function of class Trainer, we need to be decided whether the training should be conducted by words or n-grams. In the case of n-grams, it is necessary to rebuild the training set on the basis of variations with repetition of words. Training of individual words and bigrams is very fast. There is an almost linear[32] CPU and memory requirements due to the number of words in the file used for training. The increasing size n, when n-word is used, leads to CPU load and memory requirements are exponentially dependent.

Subsequently, the relative frequency of words or n-grams in the training set is calculated. To facilitate this step, the function of the Trainer class can be used. This provides a base for finding the model meets the principle of maximum entropy. Using the Trainer class we ensure optimized way to store the training set to computer memory as possible in object-oriented programming.

4.4.3 Implementation of the classification process

To use the algorithm, minimal user interaction is required. He should enter or import the sentences, which should be rated by the degree of sentiment. Sentences are separately tokenized, normalized and then passed to the Model class. Within the class of distribution functions are calculated for each word in each sentence separately, using the conditions to generate models and GIS optimization algorithm is looking for the best fitting model. It

satisfies the condition of maximum entropy.

Based on the obtained model, empirical probability distribution is designed and the probability of belonging to a possible classification classes is estimated. The probability of belonging to the sentimental sentences is of interesting for us in this case. Labor that can be termed as Sentiment score (shorter Sns).

4.4.4 Software output explanation

Output algorithm parameter (Sentiment score) always takes values from 0 to 1. Zero should be interpreted as completely neutral sentence. A property of the algorithm described above is the fact that a completely unknown sentences assigns a value of 0.5. The value of Sentiment score of 0.5 should be interpreted as unclassified due to insufficient information. Logically, sentences rated more than 0.5 have a growing sentiment.

Chapter 5

Corpus and dataset

After deciding for supervised training MaxEnt classifier it was necessary to find data for training and testing. As was written in the chapter 2, we need two class of training data. Used method works best with pre-classified sentences. Classified longer texts or articles are useable, too but they are contained by both classes fragments.

So, optimal training dataset for MaxEnt is based on short simple sentences. Moreover, it's appropriate to remove all prepositions, conjunctions and other parts of text that cannot be the bearer of sentiment. It is very important is also method for data preclassification is done. The best situation is with hand classification. It can be done with strong sentimental sentences quite easily, neutral sentiment sentences preclassification is a bigger problem. Note there isn't actually really quality corpus for learning neutral sentiment classifier.

5.1 Existing corpora

Of course, I have tried to find the best dataset fitting to our problem. I have checked several free available corpora. They were often created to evaluate the satisfaction of customers or users. So, they distinguish positive and negative sentiment class. Corpus with neutral classified sentences is unique. Originally, I hoped to succeed combined corpora in order to generate a sufficiently large set. As a result, we can have an almost infinite set of sentimental sentences. Many studies are pointing to the results based on twitter data mining.

5.1.1 Sentiment140 twitter corpus

First of twitter based corpus is published by Sentiment140[2]. It is data set of data obtained from automated processing tweets. More specifically, the tweets from which were removed emoticon.

The corpus includes around two hundred thousand twitter post, categorized by sentiment polarity. Although, authors proclaim there are three class of sentiment object - positive, negative and neutral it's not true for currently version. So, we are only able to select derived training and testing dataset from sentiment positive+negative and neutral. Note that this corpus contains slang and other artifacts that are not commonly found in standard form of English.

The corpus is available only in English, distributed as csv file from Stanford webpages[21].

5.1.2 MLSA German corpus

Multi-Layered reference corpus for German Sentiment Analysis (shorter known as MLSA) is manually processed corpus of German sentences. They are cataloged into categories, which were primarily monitored objectivity and subjectivity. The creators of these characteristics describe the three values - negative, positive, neutral (here referred to as empty). They are still assigned to two flags - intensifier or diminish.

The disadvantage of this corpus is small comprehensiveness. It includes only 280 sentences of the German language. If we want to use the corpus to distinguish sentences including sentiment and neutral sentences, there is a training set with less than thirty valid members. So, I decided to reject of using this corpus.

The corpus is available from the web[22] site under the section of the company AKSW.

5.1.3 MPQA opinion Corpus

Data is automation categorized, but the latest version of corpus is hand filtered. So, there aren't incorrect sentences. The corpus is consisted from twenty thousand subjective expressions. That's divided into 535 documents, totally there are over ten thousand English sentences. Division of corpus is little difficult because dataset includes classes like Neutral, Positive, Negative and Both. What is Both type emotion is virtually unclear.

For goal of this thesis, we could take neutral class and opposite class composed from positive and negative class. This corpus is suitable for large contexts because it is classified by articles not by sentences.

The MPQA corpus is available from website and corpus documentation should be found in T. Wilson [1] thesis.

5.1.4 The MICRO-WNOP Corpus

It's synset oriented database. Can be used as n-gram source but isn't full useable for our project. Synsets are handy classified and they have positive or negative sentiment probability. Zero probability for both implicates full neutrality and objectivity for n-gram. Corpus is available from official web pages [11].

As source of MICRO-WNOP Corpus was used General Inquirer lexicon. There are total 11788 terms. The number of Positive is 1915 and 2291 are labeled as Negative. The remaining 7582 terms, not belonging either to Positive or Negative, can be considered to be (implicitly) labeled as Objective. Objective classification isn't much reliable which is evident even after a cursory inspection data in the corpus.

5.2 Creating new corpora

After examining the available corpora, I came to several conclusions. There is probably nobody who created corpus directly in order to confirm or refuse the presence of sentiment in text. Moreover, if we look at the individual data contained in these corpora, that do not meet the needs of the training set of a classifier, which is needed.

Although manually rated the training set are almost perfect support for classifier quality, we can see that interesting results can be achieved in automatic data mining. Automatic processing of training sets usually contain fragments of misleading information, but when properly select data sources and filtering, it may also work.

The big advantage is the robustness of such sets. Data mined training set can easily cover almost the entire vocabulary and eliminate the risk of sentences in the test set, which do not carry any information about classification.

Since I have found the available corpora for our purposes as unsatisfactory, I have decided to developing method of finding the my own corpus.

5.2.1 Multilingual corpus needs

From the beginning it's searched for such a corpus, which could be used globally. This isn't a trivial problem, it isn't easy to find data sources that exist in all languages, or at least the most commonly used.

My idea is based on observations of how large corpora were created and my previous journalistic experience. For classifier it's needed only two categories of sentences: carrying any sentiment and neutrality.

So, I designed and tested data mining on the title and introduction (perexes) tabloid magazines. At first this data was processed, and then automatically presented to the user for subjective evaluation. The advantage of this solution is the existence of tabloid magazines across languages and countries. They exist anywhere from USA to Pakistan and in each country. They are primarily reflecting the kind of emotions, on which local people are sensitive.

Another problem is ensuring multilingual neutral source text. If we take into account the possible definitions of a neutral text of the introduction chapter, we can determine that such text may not be too interesting and should be objective.

Logical and interesting option may be Wikipedia. As a neutral source text can be successfully selected articles on general topics that are apolitical, don't deal with religion, or other matters con individualist.

I tried to find information about these ideas and their realizations. But I haven't found anything. So, it looks like totally new idea to data mined gossip magazine and Wikipedia for sentiment analysis. I was very keen on testing this idea.

5.2.2 Data mining

I created a user application that allows you to process bulk data from optional sources. As resources are primarily selected web portals tabloid magazines and Wikipedia articles. It represents two sets of data, which we can assume that it was strictly designed to induce emotions or neutral submission information.

The actual data acquisition in the case of tabloid magazines is carried through the RSS 2.0 reader. It's globally independent method how we are able to download content of online tabloid magazines.

Wikipedia does not provide RSS reader or other format for bulk data download. The data collector depends on the direct analysis of html code specific wiki page. Note that all of Wikipedia pages usually aren't validly programmed and they needn't follow the same layout tags. Data collector therefore receives a sentence of CSS class mw-content-ltr.

It means the text mining efficiency is lower than the text visible on the Wikipedia pages. On some pages may carry zero if the tag is missing completely. Fortunately, we have a lot of Wikipedia pages, so big efficiency of data mining is not necessary.

Chapter 6

Metrics used in the project

Along with the development of classifiers question is arising how these classifiers qualitatively assessed. Here we are dependent on statistical methods, standard tests and graphical interpretation leading to a rapid understanding.

Parker[25] in his article presents the basic tools for evaluating the performance of classifiers. It led me to statistical theory. I combined gained experience with other sources and I used it to obtain how well different combinations of classifier settings and corpora work. Please, let me introduce variable marking used in this chapter.

- vectors of real values, each vector associated with a label (0 or 1) where we call 0 the positive label and 1 the negative label
- H is classifier
- R is output real values of classifier (it could be called score)
- T - test dataset of n labeled vectors
- S - vector of classifier outputs corresponding to the classifier prediction on some instance within the test set
- $f_0(s)$ probability distribution of scores positive instances
- $F_0(s)$ positive associated cumulative distribution function
- $f_1(s)$ probability distribution of scores negative instances
- $F_1(s)$ negative associated cumulative distribution function
- π_0 empirical probabilities positive classes
- π_1 empirical probabilities negative classes
- $f(s) = \pi_0 f_0(s) + \pi_1 f_1(s)$ overall distribution of scores
- t threshold on distribution ($s < t$ will be predicated to be positive by classifier)

6.1 Confusion matrix

Also known as confusion or error matrix pivot table. It summarizes the frequency of objects that are classified as True Positive, False Negative, False Positive and True Negative for specific thresholding value.

The confusion matrix can be estimated simpler probabilistic classifier properties such as prevalence, sensitivity, specificity, predictive value of a positive test and a negative predictive value of the test.

There is theoretical formula to calculate the above properties. The formula requires knowledge of the probability of values their classification into a particular field confusion matrix. They aren't known and therefore they are based on estimation their relative frequency.

Theoretically the specificity can be computed by equation 6.1.

$$P(\bar{A}|\bar{H}) = \frac{TN}{TP + TN} \quad (6.1)$$

It could be imagined as relative frequency of good classified negative items (or neutral sentiment).

Sensitivity value could be computed by equation 6.2.

$$P(A|H) = \frac{TP}{TP + FN} \quad (6.2)$$

It could be imagine as relative frequency of good classified positive items (or sentimental sentences).

In my case, I created confusion matrix by division corpus into two parts with volume ratio factor (1/5 to 1/3). Bigger part was used for training, the smaller for testing and computing relative frequencies. Training and testing datasets were balanced.

Finally, I tried to compute confusion matrix for hand-classified data of the same type but not from the same corpus source as training dataset.

6.2 F1 measure classifier sensitivity

One of the most easily treatable metrics is F1-measure. This method is based on a balanced harmonic mean (also called positive predictive value) and sensitivity (sometimes also called recall). Rating is calculated for a particular value of threshold.

To express the F1 equation, the introduction of new term is needed. It's precision and it may be determined from frequency of true and false positive classified samples. So, it is expressed in equation 6.3.

$$\bar{P}(A) = \frac{TP}{TP + FP} \quad (6.3)$$

The next step is putting the precision in relation to sensitivity and basic F1-measure equation 6.4 is obtained.

$$F1(t) = \frac{2\bar{P}(A)\bar{P}(A|H)}{\bar{P}(A) + \bar{P}(A|H)} = \frac{2TP}{2TP + FN + FP} \quad (6.4)$$

The formula expresses the reciprocal of the average of the inverted values. The obtained results can be interpreted as a measuring of inaccurate classification of the tested classifiers.

6.3 ANOVA

Best practice in the subjective testing world is using ANOVA metrics for presentation of hypotheses. For many people ANOVA is blackbox statistic method which says if their hypothesis should be confirmed. From statistical point of view ANOVA is combination of the methods for variance analysis. ANOVA includes two basic statistical tests. There are known as the P-test and the F-test [27]. I've used a variant of single factor ANOVA in this thesis because it corresponds to a comparison of two variance groups.

There may be a problem with ANOVA because it has three basic assumptions. Observations over data must be independent, distribution must be normal and both of the comparable variances must be homoscedalic. The last one is problem for our project, because homoscedalicity virtually means the variances must come from the same source or they must be of the same quantity.

6.4 Matthews Correlation Coefficient

Another method is well described in [26]. It takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. Other commonly used quality metrics binary classifier such conditions fail.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (6.5)$$

Note that Matthews correlation coefficient (expressed by equation 6.5) is sometimes called the phi coefficient.

6.5 ROC curves, AUC

ROC curve is one of the most used tools for qualitative visualization activities classifier. Its construction is described in detail in an article by David MW Power[26].

Formally, there are introduced two distribution functions (equation 6.6 and equation 6.7).

$$F_0(x) = P(X \leq x|D = 0) = x - \int_0^x f_0(t)dt \quad (6.6)$$

$$F_1(x) = P(X \leq x | D = 1) = x - \int f_1(t) dt \quad (6.7)$$

In practical examples, we often consider that continuous random variable is not possible. There are quite typical mixtures or discrete random variables represented in the computer.

ROC curve can then be determined by estimation of specificity and sensitivity. This is done according to relative frequency of correctly classified positive elements and false classified negative elements from testing dataset. Relative frequencies are calculated for a sufficient amount thresholds in the interval from 0 to 1.

ROC curve should be interpreted as an indicator of the performance of classifiers. Its course tells us how our assumptions used for the classification are correct.

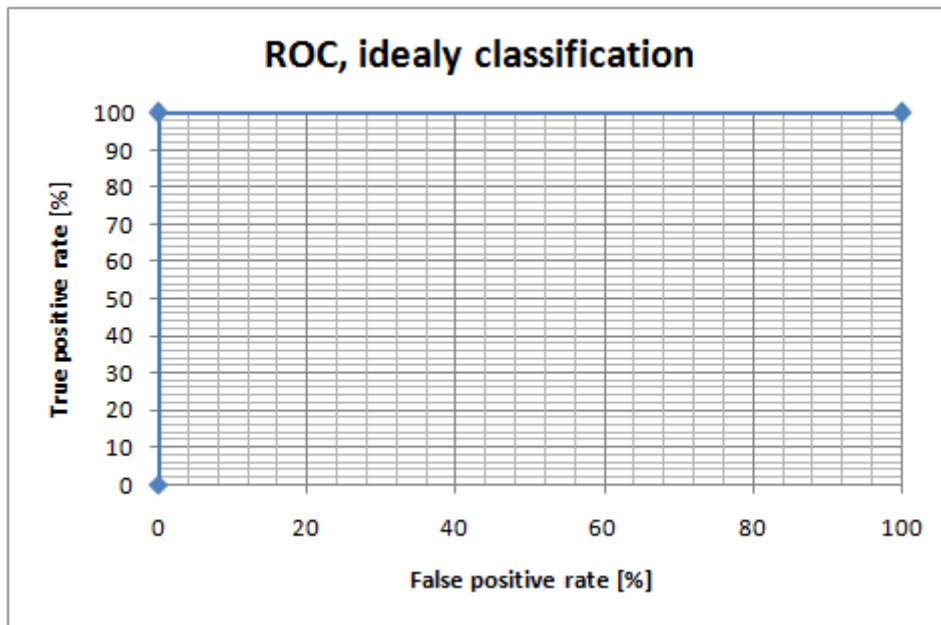


Figure 6.1: ROC ideal classified

For quick reference, here are the possible cases ROC curves. In the figure [6.1] we can see ideally ROC curve. It should significance absolutely successful classification. However, in real terms it's virtually unattainable.

When we came with wrong assumptions of classification, it may have led to two type of ROC. If assumption is totally independent to real classification symptoms, classifier works pseudo-random. ROC then looks like curve in the figure 6.2. The second type of error leads to inverse mode of classifier. This occurs when our assumptions are misinterpreted but they aren't independent.

Misinterpreted assumption [6.3] with inverse mode may be resolved by swapping classification datasets. It is always necessary to consider whether there is positive logical sequence between

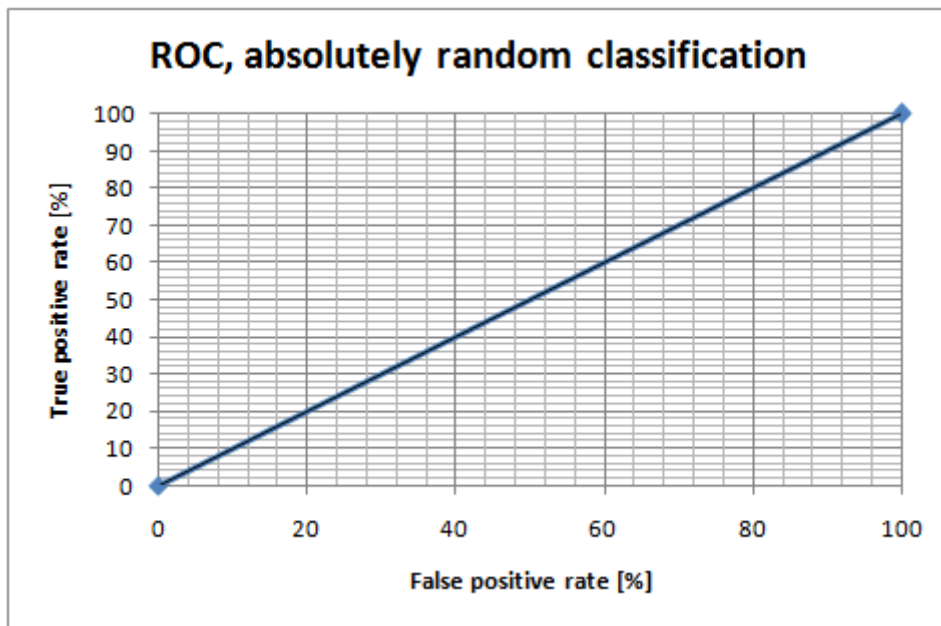


Figure 6.2: ROC random classified

symptoms and classification classes.

One disadvantage of the ROC curves is that the ROC isn't designed to use the second threshold value. It cannot display values that were excluded from the classification because they are unknown. In ex.: classifier training set is not bound by any contextual information to this values.

ROC often simplifies to a scalar value known as the AUC. Area under (RO) curve summarizes the estimate of the probability that any object on the input classifier will be classified into the correct class.

6.6 Crossvalue validation

If there is a sufficient amount of pre-classified data, we can evaluate the influence of the choice of training data on the performance of classifiers. It helps to avoid overfitting classifier and monitor the consistency of the relationship between the training data and the data which we want to classify.

The original method was designed primarily for validation of predictive models. It is also known under the name of rotation estimation. There are minimal three standard versions of crossvalue validation. I have read their methodology and I chose the most suitable for my classifier.

We can choose from K-fold cross-validation, Repeated random sub-sampling validation and Leave-one-out cross-validation.

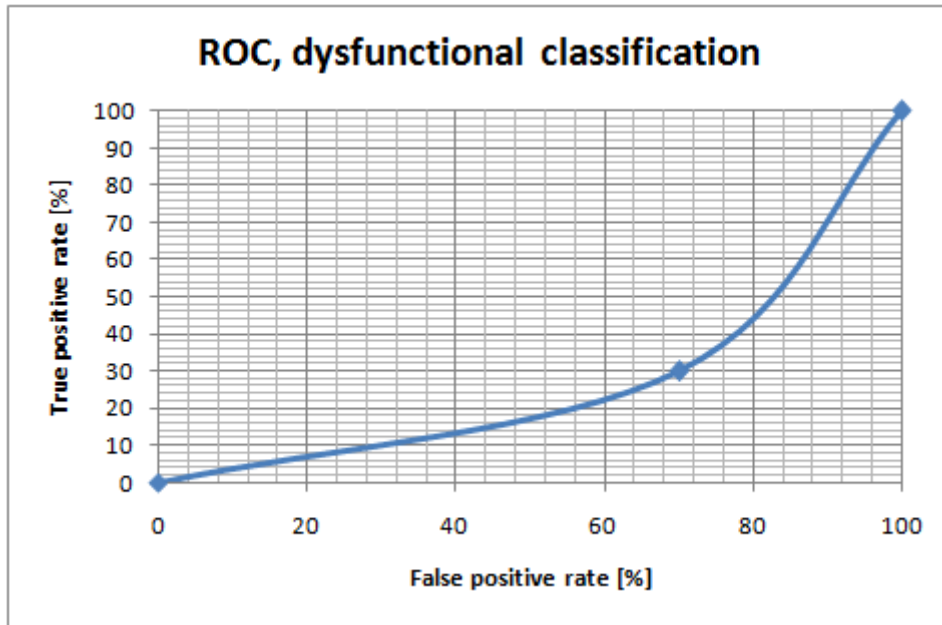


Figure 6.3: ROC dysfunctional classified

First of them, k-fold cross validation, is based on random partition of original training values. Firstly data is partitioned to k-blocks of data with equal sizes. From this k-sized blocks are k-1 vales used for training and last one is used for testing dataset. Testing value estimation is saved and the next step is data rotation. Parameter k is incremented and algorithm is repeated (originally called as foddered). After k-repeation of algorithm, every value is used for testing and training. Practice example with ten-fold cross validation can be found in the article [24].

There is the lighter version derived from k-fold cross-validation, known as hold out cross validation. At start we only randomly split values from pre-classified dataset to two classes. It's necessary to split data to finally class equal sizes. Evaluation is provided by training on the first and testing on the second data class and vice versa, respectively. There are two advantages because resultant classes are larger beside k-fold cross validation. It's pretty easy for implementation.

Compared with k-fold cross validation, Repeated random sub-sampling validation method brings an element of randomness. The first step is data splitting to two classes full randomly [18].

No matter which method we use, we observe how the results are dependent on the particular choice of training data. If the success of classifiers is very different in different folds, we face large data inconsistency. In practice, the classifier may unexpectedly fail on some data.

Chapter 7

Project results

7.1 Training corpora

During working on this thesis I have found there are many corpora focused on sentimental analysis. Only small number from them is working with neutral sentiment class. So, I decided to create myself corpus as necessary for testing my hypothesis. Firstly I created corpus in my native language - Czech. For verification multilingualism, I moved to English language. So, I have also created English corpus.

7.1.1 Czech corpus

Finally, Czech corpus for sentiment analysis in the text is constituted from two balanced classes - sentimental and neutral text. Classes include sentence fragments. These fragments were automatic downloaded, but handily filtered and combed. There are only simple sentences in corpus. Some composed sentences were divided and stored only for parts, which bearing sentiment information.

Used sentence fragments have been obtained namely from online magazines Blesk, Prásk and Super. As neutral class, Czech version of Wikipedia sources were used.

Corpus has actually 12854 sentence fragments. Corpus doesn't contain sentence punctuation characters.

7.1.2 English corpus

English text sentiment analysis is performed on combination two types of data. Neutral class is more reliable because there is used pre-subjective tested dataset. As source was used strong data provided by combination open twitter corpora. I tried to maximize using available neutral data. So, especially neutral class was revised manually. Sentimental class training data was taken directly from existing corpora and it was only subjected to minimal revision.

For better understanding, it's good to know how sentimental data was automatically selected. Standard twitter messages (tweets) were progressed by emoticons. Corpus is contained by

slang, too. This isn't problem when we keep formal grammar on classifier input. Subjective testing use just formal grammar.

Classes which English corpus is consisted are prepared as balanced with same mean as in Czech corpus. Actually, there is 2448 sentences and corpus doesn't contain sentence punctuation characters.

Note, I tried to apply idea with Wikipedia and gossip magazines too. But in the case British and USA online gossip magazines I found articles are too monotonous. It led to small corpus diversity. So, I had no choice and I had to reject this method of sentiment mining.

7.2 Project metrics fulfillment

7.2.1 Classifier confusion matrix

The confusion matrix was computed for normal classifier treshold 0.5. It doesn't shown all reality, because confusion matrix lacks number of sentences which can't be classified. Totally count of non-classified sentences is six. Totally classified sentences is 190.

		Original class prediction	
		Neutral	Sentimental
MaxEnt classification	Neutral	95	2
	Sentimental	3	88

Table 7.1: Confusion matrix of sentiment classifier

There are three sentimental sentences which are classified as neutral and only two neutral sentences classified as sentimental. This is accuracy over 93 %.

7.2.2 ROC analyses, AUC

To verify the classifier, the properties are well obtained from ROC analysis. Ready classifier implementation was verified by dividing the corpus into training and test set. The ratio of training data and testing is 5:1.

The test included 786 training sentences and 196 testing sentences. Iterative thresholding was processed by selection in increments of 0.01. The resulting ROC curve we can see in the figure 7.1 below.

The area under the ROC curve (AUC) covers 89 % of the area. It doesn't mean that the classifier ranked 11 % of sentences in the wrong class sentiment. Six sentences were excluded from the classification because the classifier has not been adequately trained enough in them to find the information needed to complete the classification. These sentences are labeled as unclassifiable on output.

Note, the number of non-classified sentences can be reduced for future by increasing size of training corpus.

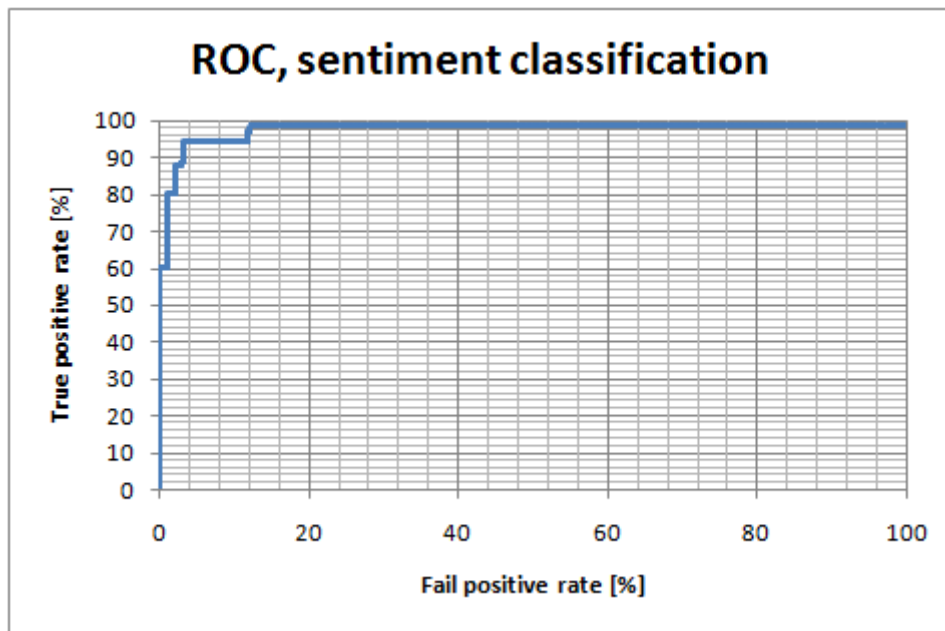


Figure 7.1: ROC of implemented classifier

ROC analysis shown the best threshold is about 0,624 when classifier put only five sentences to bad class. But this threshold leads to overfitting. It's highly recommended to stand two other treshold upper and lower then optimal for selecting data with greater reliability.

7.3 Subjective measurement data and classifier results

It was able to obtain data from a subjective test of Audience Inc. These data include the next test of English phrases also relevant MOS parameters and their variances. There were a total of 90 tested phrases, which are of two thirds simple sentences and the remaining third sentences are composed.

After verifying the functionality MaxEnt classifier I trained classifier on my own English corpus. Individual test phrases Audience Inc. were classified. There were associated with the probability of belonging to the sentimental class. Examples of such evaluation can be found in the following table.

Five sentences best fit to neutral class:

- The bark of the pine tree was shiny and dark (0,064)
- The pennant waved when the wind blew (0,102)
- The hog crawled under the high fence (0,143)
- Burn peat after the logs give out (0,182)
- The grass curled around the fence post (0,290)

Five sentences best fit to sentimental class:

- This has been a pretty good quarter yeah our sales have been a lot better than expected. (0,980)
- This is been a pretty good quarter yeah our sales have been a lot better than we expected. Someone should clean that up. (0,973)
- Excuse me; do you know when the next shuttle will be leaving? I don't want to miss my flight. (0,891)
- The new girl was fired today at noon (0,869)
- Attention please, the departure gate for Flight 2345 has been changed from 26A to 30B. (0,859)

7.3.1 Correlation

As was stated at begininng of this thesis, our hypothesis is the variance of MOS (vMOS) is dependent on Sentiment score (4.4.4) in test sentences (Sns). We expected A-factor of E-model (Expectation factor) has sentiment as significant component.

From equation emodel we can obtain the dependency should be linear. So, correlation should be one of useful metrics. I tried to compute correlation coefficient (Person's corelation coefficient) for tested sentences. There is modified equation 7.1 for computing correlation coefficient between MOS and Sentiment score. Sns is Sentiment score and vMos is variance of MOS parameter.

$$\rho_{(vMOS, Sns)} = \frac{E(vMOS, Sns) - E(vMOS)E(Sns)}{\sqrt{E(vMOS^2) - E^2(vMOS)}\sqrt{E(Sns^2) - E^2(Sns)}} \quad (7.1)$$

Corelation coefficient over all testing data is 0.104. It can be concluded that there is a positive correlation between sentiment in the test sentences and endpoint MOS. It is also clear the influence sentiment isn't prevailed in the measurement MOS result.

7.3.2 ANOVA single Factor

Groups	Count	Sum	Average	Variance
vMOS	450	443.4336347	0.985	0.014
Sns 4.4.4	450	254.957	0.566571444	0.032

Table 7.2: ANOVA summary for MOS variance and Sentiment score

According to section 6.3, I tried to make simple analysis of sentences sentiment rating and their MOS respectively. In this case, our hypothesis is the MOS variation depends on sentiment in sentence. It leads to results of statistical F-test and P-test. For F-test was counted output value 1699.077 while 3.851 is critical value. P-test output value is very close

to zero.

P-test and F-test results may be interpreted as very strong presumption against neutral hypothesis. Although this corresponds to our observation, there may be an error of the first type (false rejection of the correct null hypothesis). Please keep the fact, our testing data are limited and still there is the possibility that will be found some form of addiction in the future.

Source of variability	Sum Sq	df	Mean Sq	F val	P val	F crit
Between Groups	39.470	1	39.470	1699.077	2.7E-209	3.851
Within Groups	20.860	898	0.023230504			
Total	60.331	899				

Table 7.3: ANOVA results for MOS variance and Sentiment score

We can obtain there are only two from three constraints well satisfied. It's not clear if we are able to satisfy homoscedascity constraint. We cannot say anything about sentimentality variance. So, ANOVA may not be meaningful in this case.

7.3.3 Live testing data behavior

In this thesis, we were very strictly limited to industrial data from a single company. Subjective testing sets high personal and financial needs, so we could not get our own data or influence the corporate testing.

All conclusions are related to company data Audience Inc. and may not reflect the whole reality across other subjective tests. We followed the behavior of sentences with the lowest and highest sentiment. Specifically, the dependence of the MOS and the sentiment scattering is observed.

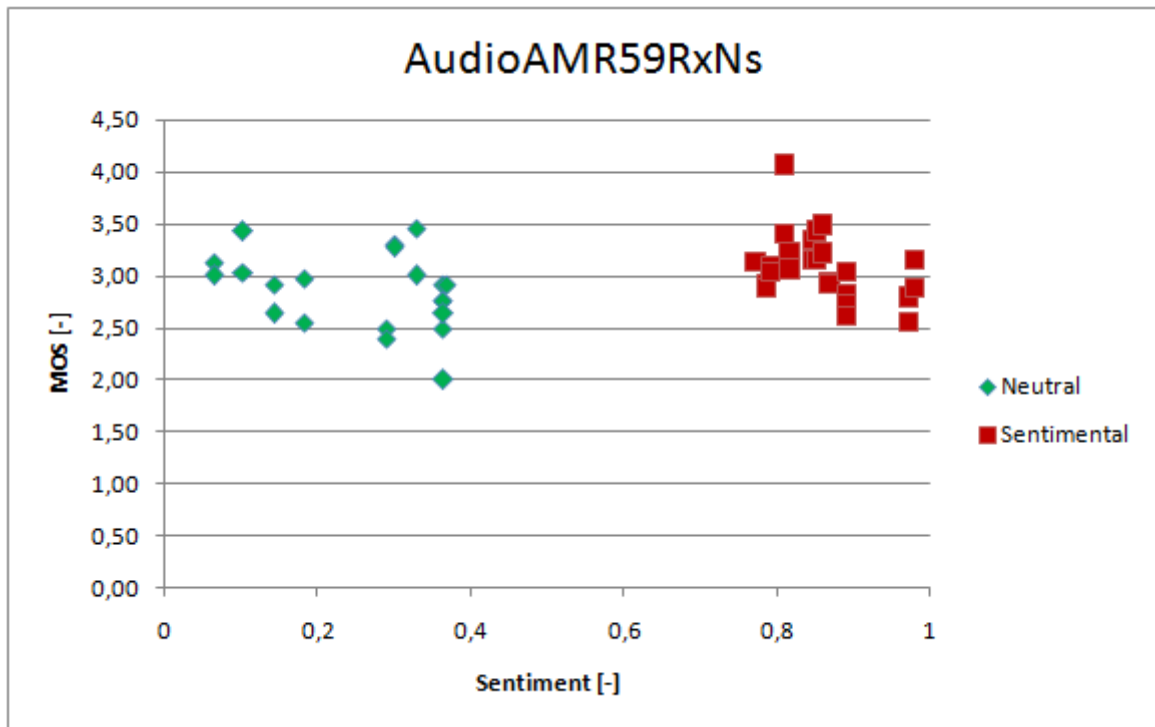


Figure 7.2: MOS score for neutral and sentimental sentences, very low quality network

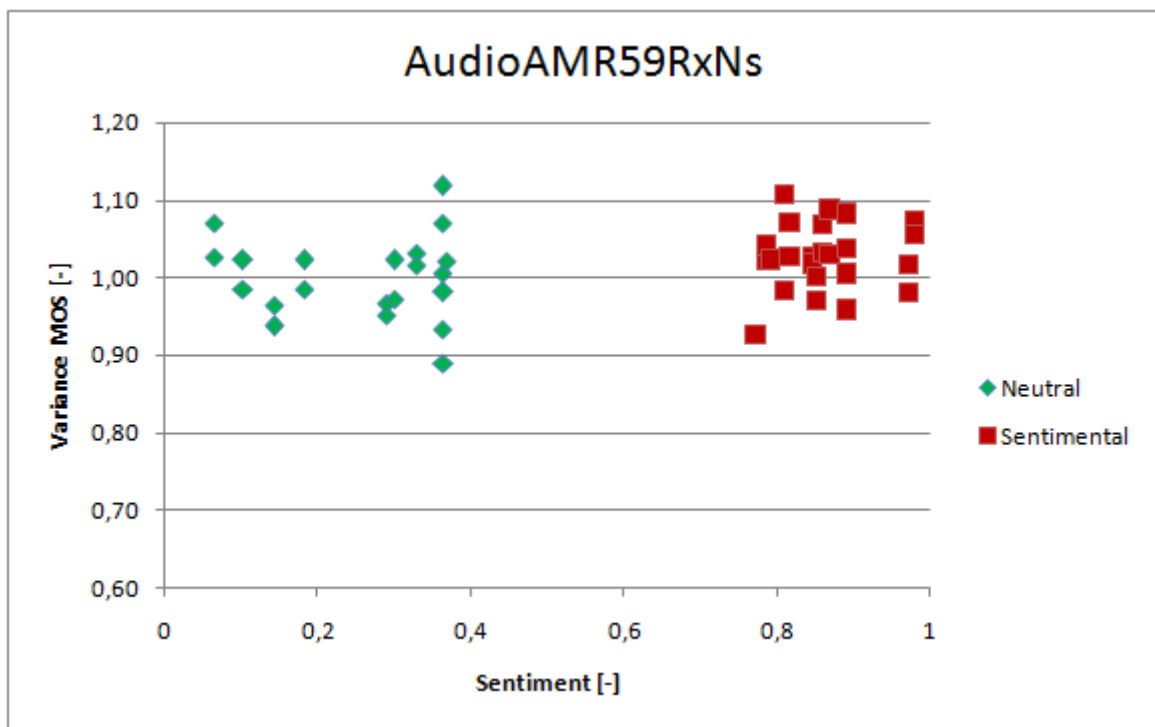


Figure 7.3: Variance of MOS score for neutral and sentimental sentences, very low quality network

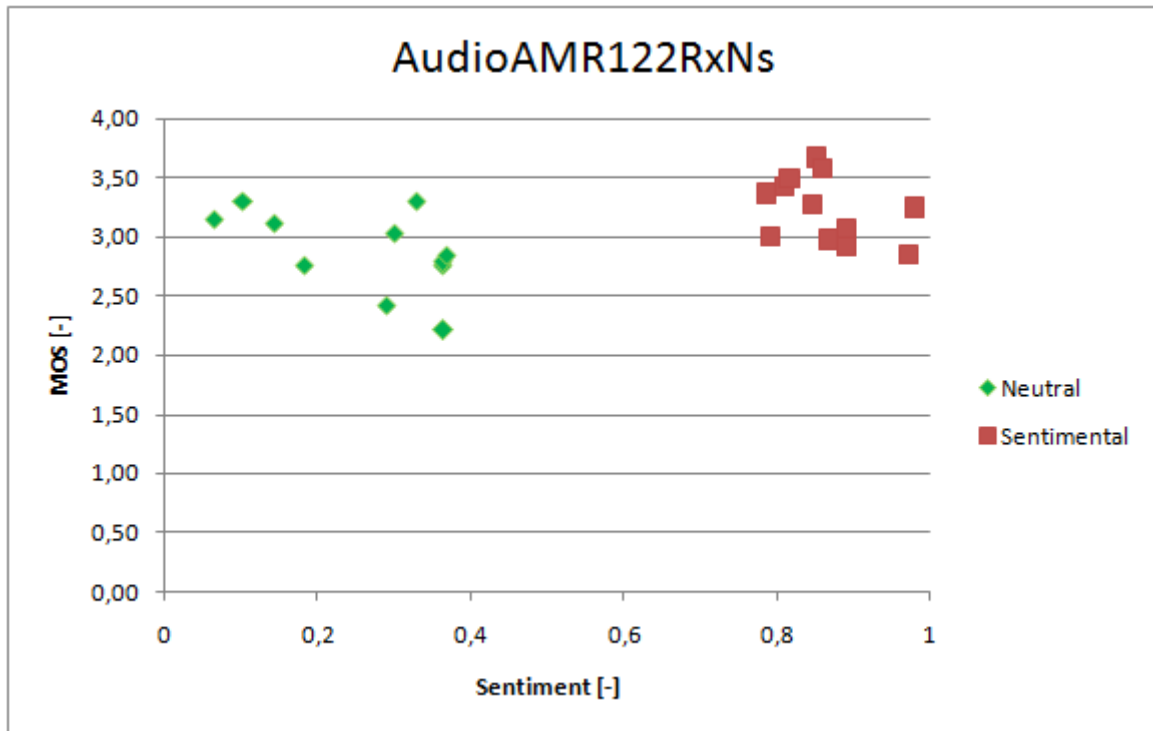


Figure 7.4: MOS score for neutral and sentimental sentences, low quality network

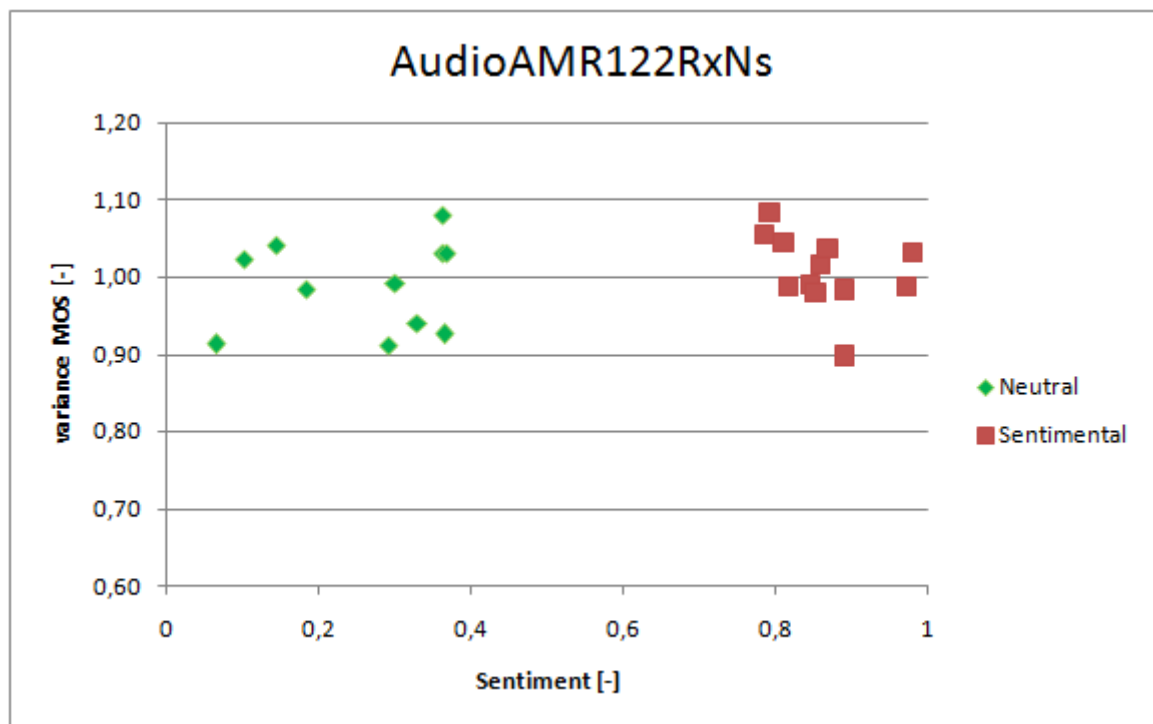


Figure 7.5: Variance of MOS score for neutral and sentimental sentences, low quality network

We can see in the graphs above, sentimental phrases results (MOS, variance MOS) are well camouflaged by neutral sentences results. In terms of statistics suggests that sentimental phrases really behave slightly differently from neutral. We cannot declare that the mean value of the MOS was higher for sentimental than neutral.

Note, the variance also behaves very stochastically. If we start from scattering throughout the test, we find that the normal probability distribution. Extremely neutral and sentimental sentences have approximately the same variance.

It must be said that the test data from the Audience Inc. does not fully correspond with the form of sentences that I expected. We hypothesized; subjective quality of voice transfer testing uses appropriately selection sentences of ordinary human communication. Instead of this, it uses Audience Inc. used to test sentences from fiction literature. This causes inconsistency between our classifier and the results of Audience Inc. testing. However, we can exclude significant number of serious errors in classification. The results are in good agreement with reality as far as human can judge.

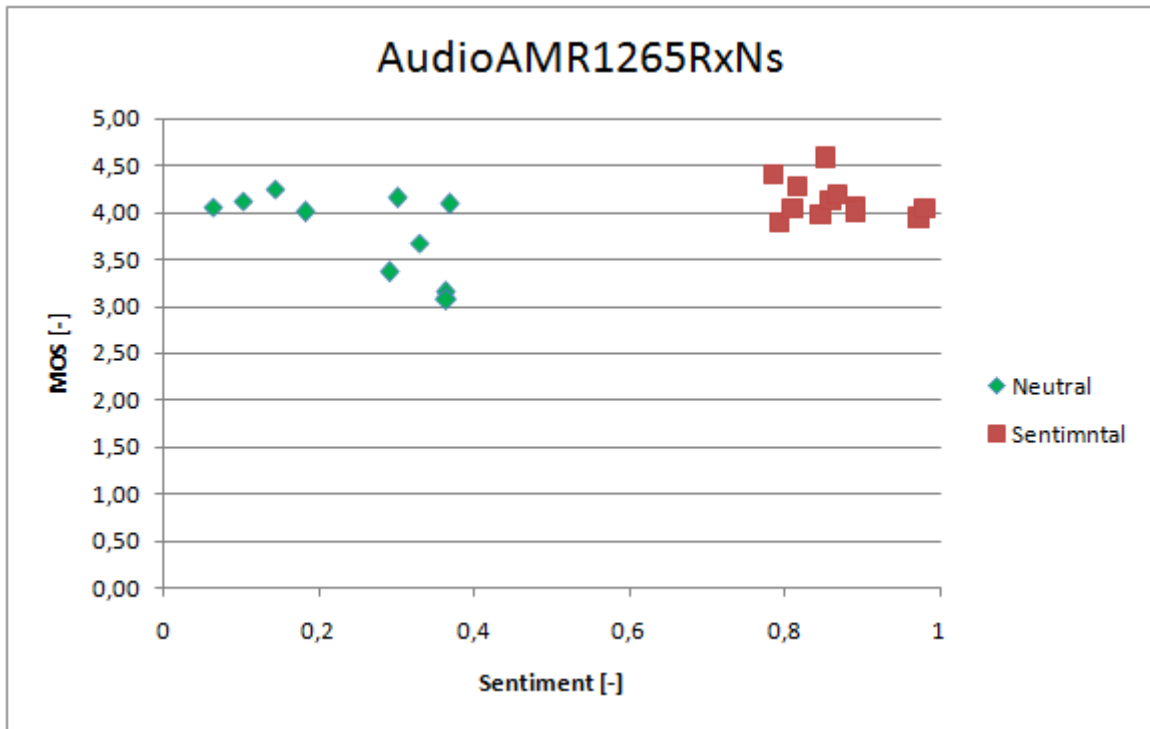


Figure 7.6: MOS score for neutral and sentimental sentences, medium quality network

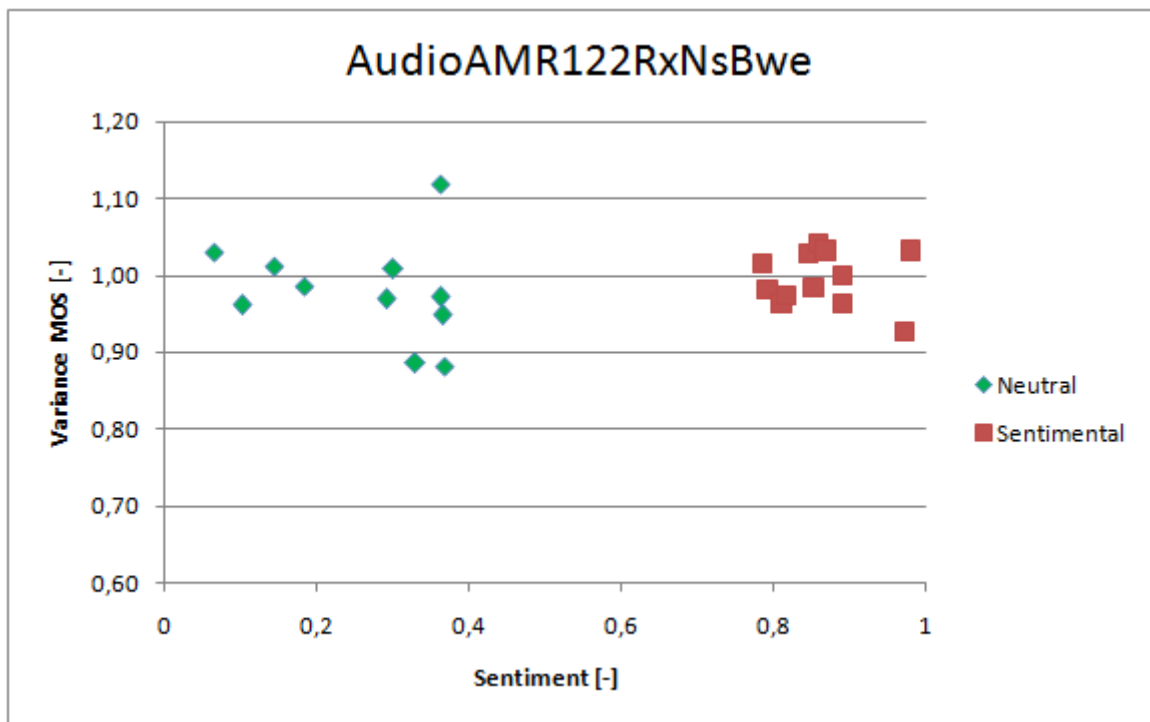


Figure 7.7: Variance of MOS score for neutral and sentimental sentences, medium quality network

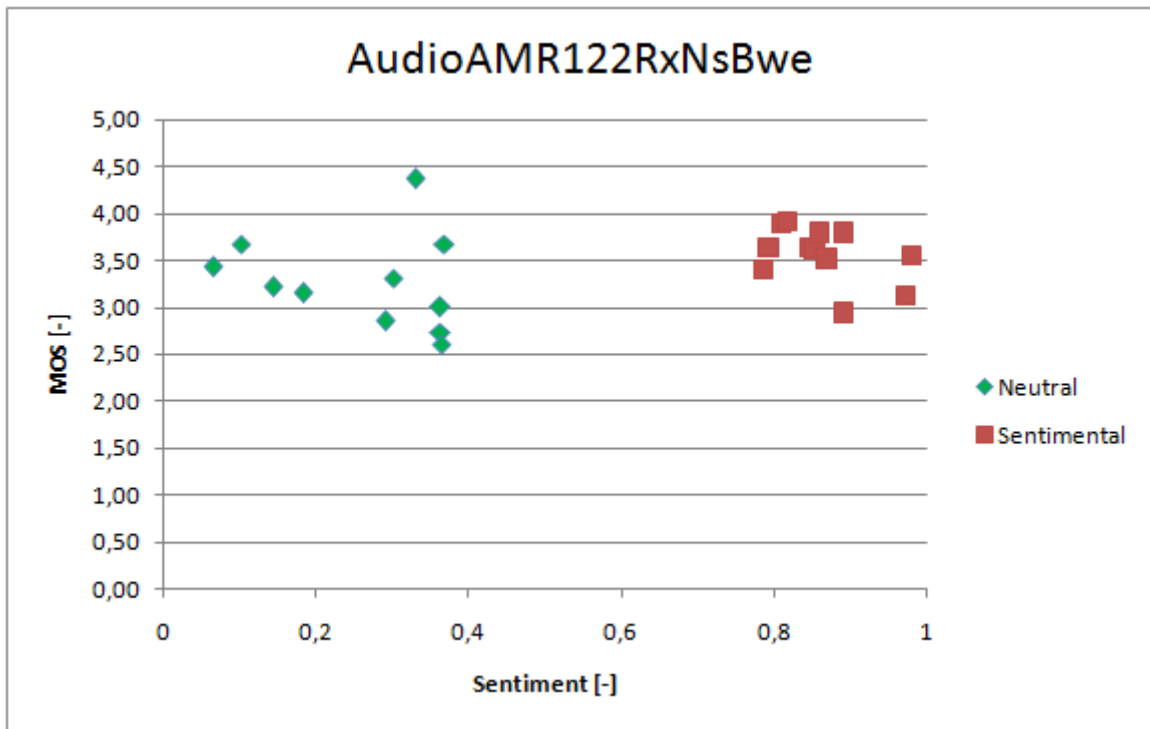


Figure 7.8: MOS score for neutral and sentimental sentences, upper medium quality network

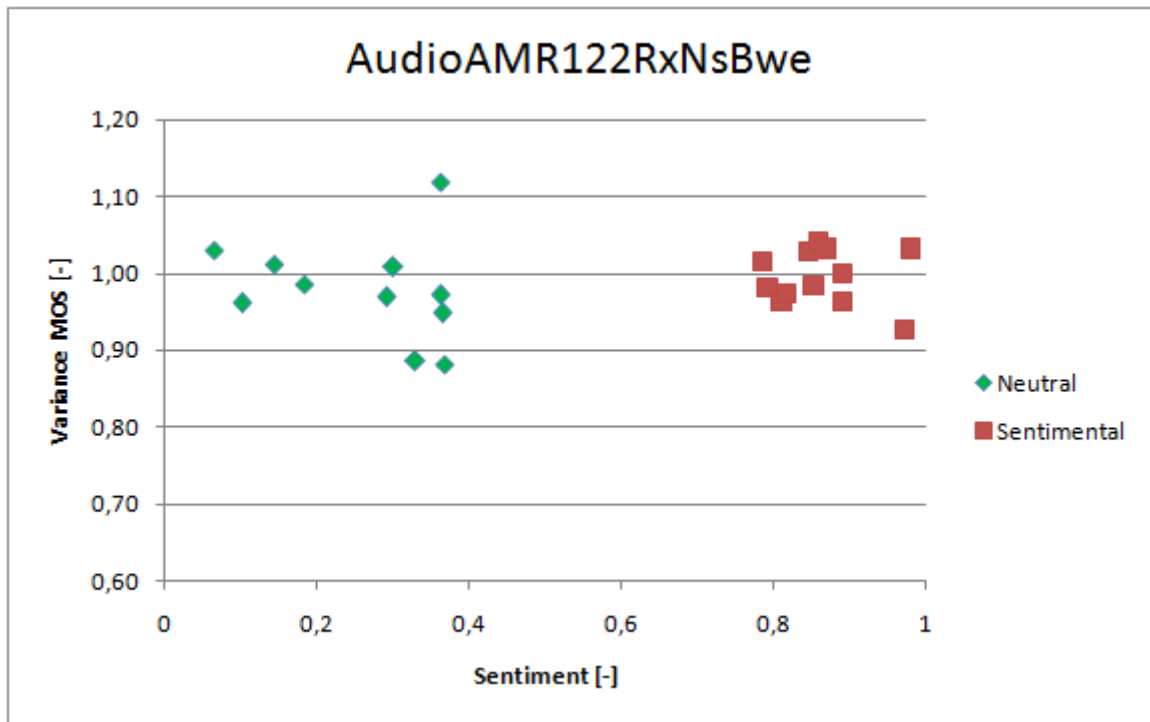


Figure 7.9: Variance of MOS score for neutral and sentimental sentences, upper medium quality network

7.4 Repeatability of results

Accompanying the subjective testing quality of voice transfer is the test conditions cannot be reproduced in full. Therefore, you cannot directly compare a couple of different subjective tests performed by P.800 recommendation. Output MOS parameter may be varying up to 50 % of its range. There are not many materials available to deal with this issue. So, we can only inspire current ideas and views on the Audience Inc. data of subjective quality tests in small.

As described in [4], there are minimal three factors which leads to MOS variation. These are cultural variation, individual variation and balance of conditions. According to the authors [4] cultural variation is significant problem. It may leads to improvement or worsening up to one full MOS point.

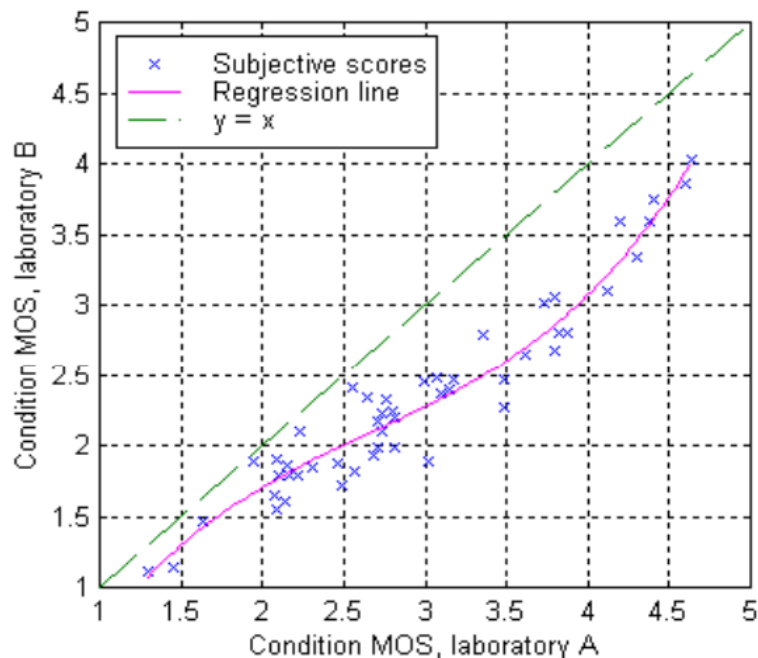


Figure 7.10: Example of MOS cultural dependency (img. source [4])

I can only confirm that cultural differences can cause problems. When processing the corpus for MaxEnt classifier, I found that the English phrases in the training set are treated differently, for example, in the USA and UK. The same data source cannot be used to select the appropriate test sentences for various cultural society. I found only one scenario direct comparison of the data of cultural different subjective tests, which can be seen in the figure 7.10. According to the source, it is not clear source laboratories. So, this scenario could be created as purposefully and pessimistic.

The second significant effect for MOS is individual variation. It may be classified like combination of self expectation and tester experience. This is in good agreement with the A factor, which is described in the second chapter as part of the E-Model. By [4] it obviously MOS is affected in range from 0.2 to 0.5.

The last main effect for MOS should be balance of conditions. This includes conditions under which it is tested and overall layout of the test. If poor samples prevails in the test, samples may be sometimes classified as bad, can be classified as good in this test. We can imagine the opposite situation, too.

There is one side effect that hasn't been described yet. Audience Inc. testing data includes repetition of two phrases in one test. The test included the same readers and listeners during testing. This can eliminate a significant change in conditions during testing.

Here are the data obtained from repeated testing of sorted according to the quality of the network settings 1 to 5 as in school according to Central European standards. PH1 phrase was tested three samples before PH2 phrase. Used phrases:

- (Sample A) Well, I hope so. Let's just go, and we'll see, right? OK, bye, see you later.
- (Sample B) Excuse me; do you know when the next shuttle will be leaving? I don't want to miss my flight.

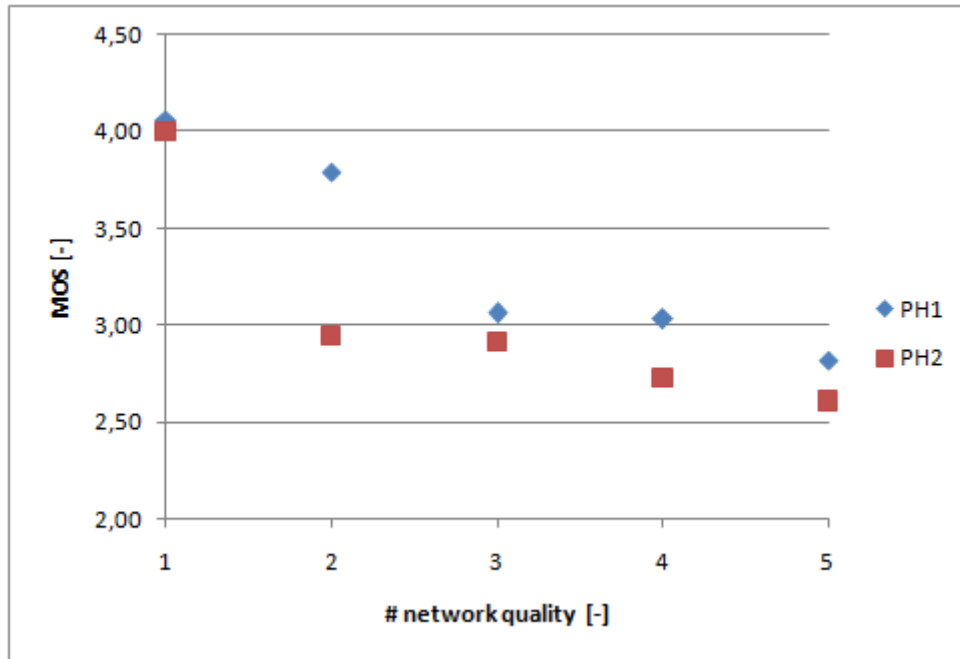


Figure 7.11: Subjective test repetition Sample A, PH1 and PH2 MOS

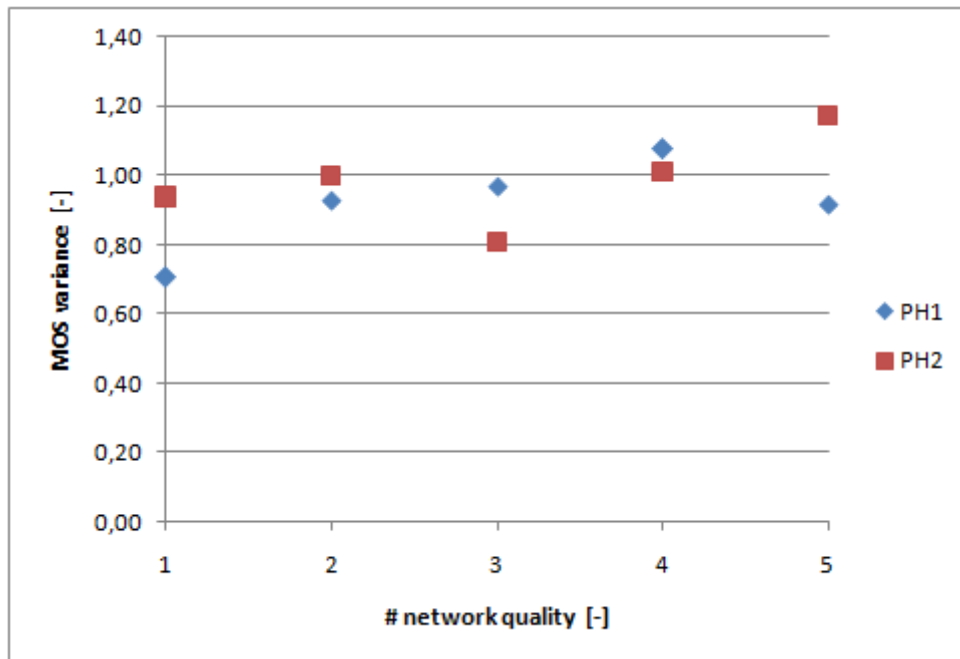


Figure 7.12: Subjective test repetition Sample A, PH1 and PH2 MOS variance

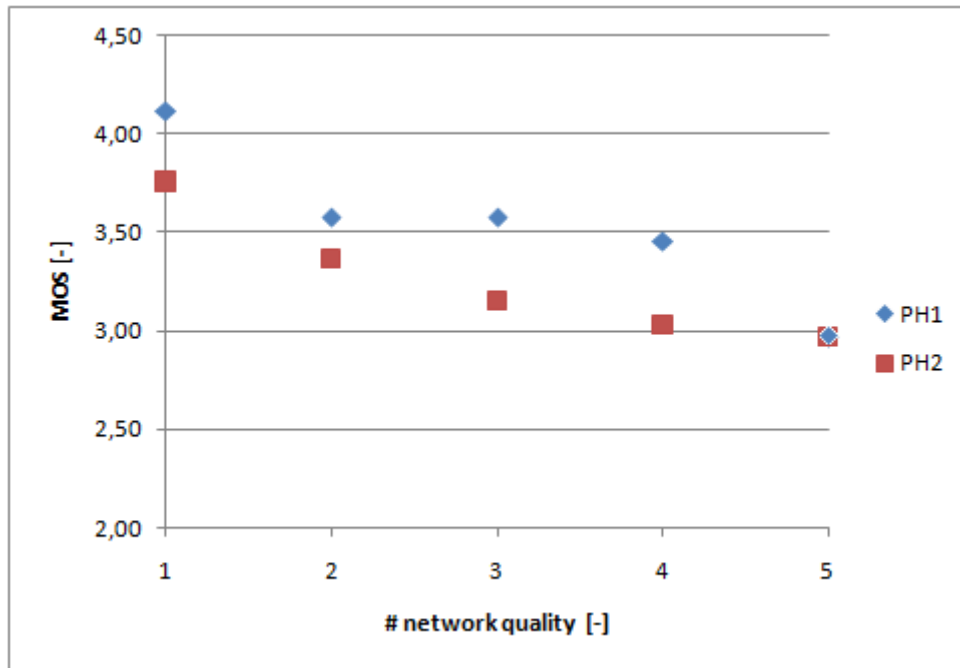


Figure 7.13: Subjective test repetition Sample B, PH1 and PH2 MOS

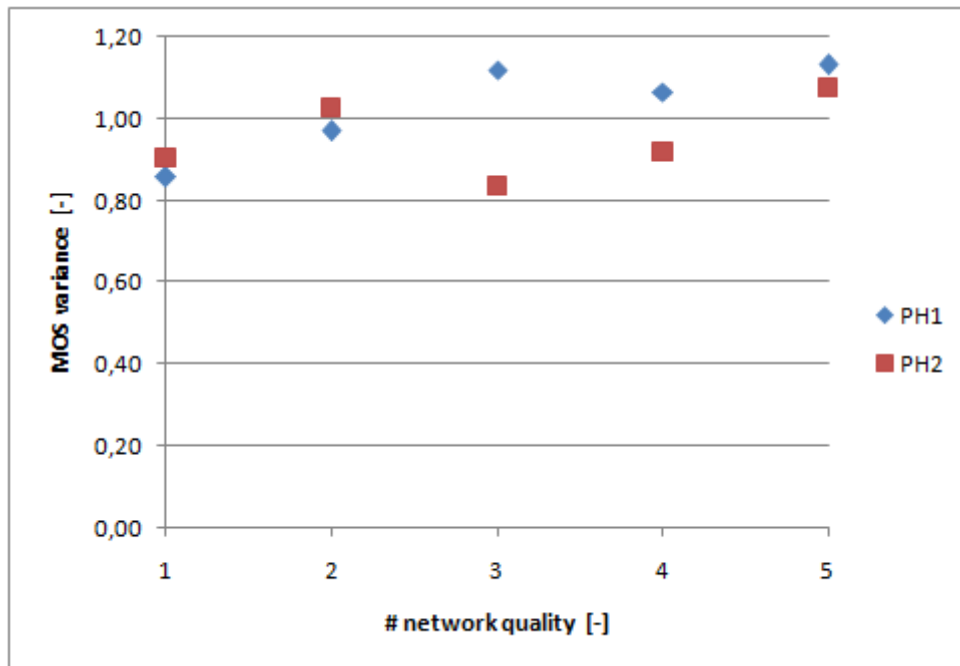


Figure 7.14: Subjective test repetition Sample B, PH1 and PH2 MOS variance

I would like to mention the sentences were not neutral. Achievements sentimental scores are 0.701 (should be imagine as sentence slightly above the threshold force for neutrality) and 0.891 (should be imagine as sentence with quite considerable sentimental component). In total we have two pairs of subjective tests on the same sentences of five different network settings. From these two observations cannot be identified serious conclusions.

However, we can observe the trend small deterioration MOS (Sample A in the figure 7.12 and Sample B in the 7.14) rating when repeating the same phrases over subjective testing. Also scattering behavior cannot be fully explained yet. Next interesting thing is there is small decrease MOS when sentences were repeated (Sample A in the figure 7.11 and Sample B in the 7.13).

It can be concluded that the repeatability of subjective testing quality of voice transfer with same conditions looks to be good. Dependence drift MOS parameter on time is negligible compared to the influences acting of changing cultural, individual or conditions balancing parameters.

Chapter 8

Conclusion

8.1 Summary

This diploma thesis combines machine learning with sentiment analysis and it puts them to context of subjective testing of voice transfer quality. It's based on official recommendation ITU-T P.800 and it examines the impact of sentiment to test results. Thesis is based on earlier developer experience, but there are new ideas and their realizations, too.

Primary target was to explore current options for development sentiment detector for typewritten text. These are texts, which are used as source for recording test samples and in next step; these samples are transferred over communication network for subjective evaluation.

During the work I have successfully defined the parameters of our problem and found adequate and functional solution. The critical point of this work was to obtain data that comes from the actual subjective testing of voice quality. Thanks to them I could compare the real impact sentiment in the results of subjective testing. Finally, I used data mining, machine learning, statistical data processing and classical object-oriented programming.

Here is one disappointment, because the initial hypothesis (sentiment leads to a significant increase in MOS variance, which is the result of subjective testing) cannot be confirmed. On the basis of data from the Audience Inc. we can reject this hypothesis. We must be restrained, because we managed to get the data from a single complete testing. It cannot be excluded, during future testing with different conditions; it will be seen noticeable improvement due to sentiment.

8.2 Benefits for future projects

The results of this thesis can be used in future projects dealing with the investigation of sentiment in the text. I created two applications for laboratories. These applications can facilitate the formation and testing new corpora or provide tools for the classification and handling of the classified data.

Thanks to the creation applications it can be automatically filtered sentences with the lowest level of sentiment and increased the credibility of the process by subjective testing

recommendations P.800. Developers of networks for voice transmission can use this thesis to determine, how sentiment contented source will effect on user satisfaction. This is especially useful in the design of conventional telephone networks, VOIP and virtual audio channels.

According to described and generated resources will be easily to switch to another type of natural language problem than the detection sentiment in the text. For example, an appropriate adjustment to the corpus would be possible to distinguish between positive and negative charge in the text or watch directly the desired type of emotions including their intensity.

8.3 Personal benefits

Work on this thesis was a great personal benefit for me. Although, I originally wanted to go to a hardware-oriented project, this theme captivated me by completely newfound topic, which required development and at initially didn't promise exact results. For me as student from Department of Measurement, the topic was particular interesting according to the need measuring the quantity which is dependent on the abstract work of the human mind that hasn't been adequately described and understood, yet.

I was working alone during the development and I made decisions about risk management and time management on my own responsibility. I was able to making free decisions in the preparation of research, the choice of methods and implementation. The operating principle of subjective testing and external influences were absolutely new for me. So, I have verified the statistical analysis of the data from these tests and I have expanded my experience with classifiers.

Bibliography

- [1] T. W. (2008). Fine-grained subjectivity analysis. PhD Dissertation, Intelligent Systems Program, University of Pittsburgh.
- [2] R. B. Alec Go and L. Huang. Twitter sentiment classification using distant supervision.
- [3] P. T. P. D. H. Andrew L. Maas, Raymond E. Daly. Learning word vectors for sentiment analysis. 142-150, ISBN: 978-1-932432-87-9 (2011).
- [4] P. L. A.W. Rix. Comparison between subjective listening quality and p.862 pesq score. September 2003.
- [5] S. D. P. Berger and V. D. Pietra. The e-model: a computational model for use in transmission planning. Recommendation G.107, versions (98 - 14).
- [6] S. D. P. Berger and V. D. Pietra. A maximum entropy approach to natural language processing. 22(1):39-71, 1996.
- [7] E. Boiy and M.-F. Moens. A machine learning approach to sentiment analysis in multilingual web texts. Inf Retrieval (2009) 12:526–558.
- [8] C. J. BURGESS. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2, 121-167, 1998.
- [9] M. M. Carlos Castillo and B. Poblete. Information credibility on twitter. Department of Computer Science, University of Chile 2011.
- [10] T. A. Carolin Strobl, Anne-Laure Boulesteix. Unbiased split selection for classification trees based on the gini index. Department of Statistics, University of Munich LMUn (2005).
- [11] C.-V. D. A. F. M. Cerini, S. and G. Gandini. Language resources and linguistic theory: Typology, second language acquisition, english linguistics (forthcoming). <http://www.unipv.it/wnop/micrownop.tgz> (2007).
- [12] K. L. L. CHAO-YING JOANNE PENG and G. M. INGERSOLL. An introduction to logistic regression analysis and reporting. Indiana University-Bloomington, EBSCO Publishing 2002.
- [13] P. R. Christopher D. Manning and H. Schütze. Introduction to information retrieval. ISBN 0521865719, Cambridge University Press. (2008).
- [14] J. R. Curran and S. Clark. Investigating gis and smoothing for maximum entropy taggers. Pages 91-98, ISBN:1-333-56789-0 (2003).

- [15] M. A. Hearst. Support vector machines, *iee intelligent system*. University of California, Berkeley 1998 Jun;62(6):18-24.
- [16] D. Hume. An enquiry concerning human understanding. Harvard Classics Volume 37, Copyright 1910 P.F. Collier Son.
- [17] E. K. Joseph Kosinski and A. Horowitz. Tron: Legacy. Popularization of artificial intelligence and cybernetics. Movie (2010).
- [18] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. Appears in the International Joint Conference on Artificial Intelligence (IJCAI), 1995.
- [19] R. Malouf. A comparison of algorithms for maximum entropy parameter estimation. Association for Computational Linguistics Stroudsburg, PA, USA 2002.
- [20] J. S. N Nayab. Disadvantages to using decision trees. online article, www.brighthubpm.com, (2/9/2011).
- [21] J. Nurnberger and T. Jr. Foroud. Sentiment140 twitter corpus data. <http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip>.
- [22] I. G. on German Sentiment Analysis (IGGSA). A multi-layered reference corpus for german sentiment analysis. <http://datahub.io/dataset/mlsa>.
- [23] R. L. P Melville, W Gryc. Sentiment analysis of blogs by combining lexical knowledge with text classification. KDD'09 Pages 1275-1284 ISBN: 978-1-60558-495-9 (2009).
- [24] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity, summarization based on minimum cuts. cap. 4, Cornell University.
- [25] C. Parker. An analysis of performance measures for binary classifiers. Data Mining (ICDM), 2011 IEEE.
- [26] D. M. W. Powers. Evaluation: From precision, recall and f-measure to roc, informedness, markedness correlation. Technical Report SIE-07-001: 2007.
- [27] prof.Ing. Mirko Navara DrSc. Pravděpodobnost a matematická statistika. Skriptum ČVUT, Praha, 1. vydání, 2007.
- [28] I. Rish. An empirical study of the naive bayes classier. T.J. Watson Research Center.
- [29] E. Shouse. Feeling, emotion, affect. Shouse, Eric.
- [30] Y. W. Songbo Tan, Xueqi Cheng and H. Xu. Adapting naive bayes to domain adaptation for sentiment analysis. Key Laboratory of Network, Institute of Computing Technology, China.
- [31] R. P. A. c. R. E. S. Steven J. Phillipsa, . Maximum entropy modeling of species geographic distributions. *Ecological Modelling* 28. March 2008, available at www.sciencedirect.com.
- [32] M. K. Tanel Alumae. Efcient estimation of maximum entropy language models with n-gram features. Institute of Cybernetics, Tallinn University of Technology, Adaptive Informatics Research Centre, Aalto University (2010).

- [33] I. T. Union. The e-model, a computational model for use in transmission planning. G.107 (12/98).
- [34] I. T. Union. Itu-i recommendation g.107 (2011) amd. 1 (06/2012)s. <http://www.itu.int/rec/T-REC-G.107-201206-S!Amd1>.
- [35] V. Vryniotis. Machine learning tutorial: The max entropy text classifier. <http://blog.datumbox.com/machine-learning-tutorial-the-max-entropy-text-classifier>.
- [36] V. Vryniotis. The multinomial logistic regression (softmax regression). Datumbox 2013-25-12.