ПРИЛОЗИ. ОДДЕЛЕНИЕ ЗА ТЕХНИЧКИ НАУКИ
CONTRIBUTIONS. SECTION OF TECHNICAL SCIENCES

Прилози, Одд. тех. науки, МАНУ	Том	Бр.	стр.	Скопје
	XXXV	1-2	1-56	2014
Contributions, Sec. Tech. Sci., MANU	Vol.	No.	pp.	Skopje

UDC 62 ISSN 1857-6656

МАКЕДОНСКА АКАДЕМИЈА НА НАУКИТЕ И УМЕТНОСТИТЕ
MACEDONIAN ACADEMY OF SCIENCES AND ARTS

ОДДЕЛЕНИЕ ЗА ТЕХНИЧКИ НАУКИ
SECTION OF TEHNICAL SCIENCES

ПРИЛОЗИ
CONTRIBUTIONS
XXXV 1-2

СКОПЈЕ – SKOPJE
2017

TAKI FITI, PRESIDENT OF THE MACEDONIAN ACADEMY OF SCIENCES AND ARTS

ON THE EVE OF THE GREAT JUBILEE – 50 YEARS OF THE MACEDONIAN ACADEMY OF SCIENCES AND ARTS 1967 – 2017

This year the Macedonian Academy of Sciences and Arts (MASA) marks and celebrates a great jubilee – 50 years of existence and work of our highest institution in the field of sciences and arts. Although on 22 February 2017 the 50th anniversary of the enactment of MASA in the Assembly of the Socialistic Republic of Macedonia was marked, and on October 10 it will be 50 years since the solemn establishment of MASA, we proudly emphasize that our roots, the roots of the Macedonian and Slavic cultural and spiritual continuity, are far back, in a time dimension which is measured in centuries. Because the mission of the Ss. Cyril and Methodius, the historical events that made Ohrid, with the famous Ohrid Literary School, already in the IX century to become the center of the Slavic educational and enlightening activity, which then spread throughout all Slavic countries, have fundamentally changed our contribution to the treasury of the European culture and civilization. And furthermore, centuries later, in the middle of the XIX century the Macedonian revival began, with a pleiad of our cultural and national activists. These processes at the beginning of the XX century resulted in the establishment of the Macedonian Scientific and Literary Fellowship in Saint Petersburg, led by Dimitrija Chupovski and Krste Petkov Misirkov, whose rich scientific, literary and cultural activities were a significant reflection of our spiritual continuity and identity, and an event that has marked the dawn of the Macedonian Academy of Sciences and Arts. This continuity will remain in the period between the two world wars, with a pleiad of artists in literature, art, music, philological, economic, legal and technical sciences. A few years after World War II, in 1949, in free Republic of Macedonia, the first state University of "Ss. Cyril and Methodius" was established, within which, in less than two decades, solid personnel resources were created which allowed rapid development of the higher education and scientific activity in our country. It was an event of great importance for the establishment of MASA as the highest institution in the field of sciences and arts.

This millennium pace and continuity in the development of art and scientific thought in our region is an indication and evidence that we are not a nation without its own roots, without its own history, without its own culture, and that the attempts to deny our identity, language, name, no matter where they come from, are residual of the Balkan anachronisms, and essentially speaking, they are absurd and retrograde.

Immediately after the establishment of MASA followed a period of rapid development, diversification and enrichment of its scientific and research activities and artistic work. Almost two decades after the establishment MASA entered the phase of its maturity and has grown and has affirmed as the fundament of the Macedonian science, language, culture and history and as one of the pillars and symbols of the statehood of the Republic of Macedonia.

Today, MASA, according to its integral concept, structure and function, has all the necessary attributes of a modern national academy of European type, and of course, performs the three basic functions typical of the European national academies: creating communication space for confrontation of different views and opinions on important issues in the field of sciences and arts, scientific and research work and advisory role.

The scientific and research activities and artistic work, in fact, constitute the core of the activity of MASA. The number of completed scientific and research projects and projects in the field of arts within MASA is impressive – more than 1,000 projects in the past 50 years. Some of these projects are long-term and are mainly related to the strategic issues of specific national interest, and significant is the number of fundamental and applied research in all fields of science and art represented in the Academy. MASA members in their scientific research increasingly incorporate the international dimension in the work – in the recent years more than 60% of the scientific papers have been published in international journals, most of which have been published in journals with impact factor; 50% of the papers that have been published in proceedings of scientific and professional meetings are related to meetings held abroad, etc. In addition, the works of our renowned writers and poets, members of MASA, are translated into foreign languages, and their work has found its place in world anthologies. Our prominent painters and sculptors of the older and the younger generation have created and create masterworks that are regularly exhibited at home and abroad. It should be particularly noted that our two research centers – Research Center for Energy and Sustainable Development and the Research Centre for Genetic Engineering and Biotechnology "Georgi D. Efremov", that have gained high reputation in the region and beyond, continue to successfully maintain the attained position. The work of the other research centers also enhances, including the newly established ones, which have begun to work on significant international scientific and research projects.

In its half-century of existence and work MASA developed a rich publishing activity. Since its establishment until today around 700 titles have been published – monographs, results of scientific projects, proceedings from scientific meetings, music releases, facsimile and jubilee publications, joint publications with other academies and scientific institutions, publications of solemn meetings, special issues of the departments of MASA etc. A special contribution to the publishing activities of MASA provides the "Trifun Kostovski" Foundation that has been existing and working for 18 years.

MASA proactively follows the changes and the new trends in the scope of the advisory function of the modern European national economies, and in that context the obligations arising from the project SAPEA - Science Advice for Policy by European Academies, initiated by the European Commission in order to intensify the cooperation of the European academies within their advisory role. Through the publication of the results of our scientific and research work, their presentation to the wider scientific and professional public in the country, to the government officials, etc., MASA participates in the policy-making in the field of sciences and arts and in the overall development of the country. The maintenance of the independence of MASA in carrying out the advisory role is our highest priority and principle.

In the recent years MASA has developed extensive international cooperation that contributes to the affirmation of the Macedonian scientific and artistic work and to the increasing of the reputation of MASA and of the Republic of Macedonia in international scale. Today, our Academy cooperates with more than 30 foreign academies and scientific societies and is a member of 7 international associations of academies. In the recent years the cooperation with the academies from the neighboring countries has been intensified, as well as with the Leibniz Society of Sciences from Berlin, and also, within the so-called Berlin process (Joint Science Conference of Western Balkans Process / Berlin Process) the cooperation with the German National Academy of Sciences – Leopoldina, with the French Academy of Sciences, the academies of Southeast Europe and others.

Due to the results achieved in its work, MASA and its members have won a number of high national and international awards. In the past 50 years, MASA has won around 90 awards and recognitions – charters, plaques, certificates of appreciation, medals and decorations from national and international scientific, educational, artistic and other institutions. Particularly, it should be noted that MASA has been awarded with the high decoration Order of the Republic of Macedonia for the contribution to the development of the scientific and research activity and artistic creativity of importance to the development and affirmation of the Macedonian science and state, which is awarded by the President of the Republic of Macedonia, as well as the prestigious Samuel Mitja Rapoport award of the Leibniz Society of Sciences from Berlin, which, for the first time, has been awarded to MASA. Today, 22 members of MASA have the status of foreign, corresponding and honorary members, as well as holders of honorary PhDs at around 60 foreign academies, scientific societies and universities.

***

The developmental trajectory of MASA unambiguously confirms that the Academy, in its 50 years of existence and work, faced with periods of heights, but also periods of descents and turbulences that are most directly linked to the situation in the Macedonian society, i.e. with crisis periods of different nature – the dissolution of the former common state (SFR Yugoslavia), problems with the recognition of the international status of the country after its independence, the embargoes and the blockades of the country in the early transition years, the internal conflict in 2001 and the political crisis in the last two-three years. In such crises and tense periods the criticism for the Academy grew – that MASA is an institution closed in itself, that MASA stays away from the current issues and developments in the country, and so on. On the one hand, it is a result of the insufficient understanding of the social role of the Academy – MASA is the highest scientific institution, where hasty reactions of columnist ‘type’, with daily political features are not characteristic. On the contrary, MASA uses facts and arguments. The basic activity of MASA, the results achieved in the scientific research and the artistic work is our identification within the national and international professional and scientific community, and beyond, within our society. On the other hand, this criticism and perception of MASA has a real basis in the fact that MASA, as opposed to the huge opus of implemented scientific and research and artistic projects still insufficiently affirms the results of its scientific and artistic production to the public. It is our weakness that we must overcome in the future. Of course, we cannot and must not "turn a blind eye" to the other weaknesses and omissions which, at least from time to time, we have faced with over the past 50 years and which we will face with in the future – insufficient scientific criticism of the events in the field of sciences and arts, insufficient resistance to political influence etc. On the contrary, in the future, we will have to clearly identify the weaknesses and the oversights in our work and to find out the right approaches to overcome them.

Today we live in a world of great science. The strong development of sciences, the new technological model based on information and communication technologies, the new wave of entrepreneurial restructuring of economies and societies, the globalization of the world economic activity, opened new perspectives to the economic growth and the development of individual countries and of the world economy as a whole. However, these processes, by their nature, are contradictory. The latest global financial and economic crisis of 2007-2009 revealed the contradictions of the globalization and the discontent of the people from it – the uneven distribution of wealth and power among individual countries, destruction of the resources and the environment worldwide, exhaustion of power of the existing technology and development models. These processes resulted in other problems – refugee and migration crises, strengthening of the regional and national protectionism despite the efforts to liberalize the international trade, fencing of the countries with walls at the beginning of the new millennium, changes in the economic and technological power and of the geo-strategic position and importance of entire regions and continents, etc. Nevertheless, one thing is a fact – societies that aspire to grow into societies and knowledge-based economies more easily deal with all the above mentioned problems, challenges and risks of the modern world. Of course, moving towards a development knowledge-based model assumes large investments of resources in education, science, research and development and in culture, simultaneously accompanied by well-conceived and devised strategies on development of these crucial areas of the human spirit and civilizational endurance. Hence, this fact, undoubtedly, emphasizes the special significance of the national academies of sciences and arts in achieving this objective.

In the recent years the Republic of Macedonia has been facing with the most difficult political and social crisis in the period after its independence. We are facing a crisis of the institutions, breach of the principles of the rule of law, the phenomenon of "captured state", a decline in the process of democratization of the society and falling behind on the road to the Euro-Atlantic integration processes. The problems that are now in the focus of our reality will require major reforms, much knowledge, energy and political will to overcome them. In this sense, and in this context, the role of MASA and of the overall scientific potential of the country in overcoming the crisis is also particularly important.

The above summarized evaluations and considerations about the development of MASA in the past 50 years, about the achievements in the realization of its basic activity, about the problems it faced and faces with, about the major challenges arising from the new age and which are determined with the changes in the international and national environment, they alone define the main priorities of our Academy in the forthcoming period:

- Our long-term goals are contained in the mission and vision of MASA as the highest institution in the field of sciences and arts. The mission of MASA is through the development of the basic functions that are characteristic for all modern national academies of European type, to give its full contribution to the inclusion of the Macedonian science and art in the modern European and world trends, and our vision is the Republic of Macedonia to become an advanced society based on science and knowledge;
- In the forthcoming years the focus of the scientific and research activity and artistic work of MASA, in cooperation with the other scientific and research institutions in the country and with government experts, will be particularly focused on the elaboration of issues and topics that are most directly related to the sources of the current political and social crisis in the country in order to offer possible solutions, approaches and policies to overcome it;

- The issues related to the Euro-Atlantic integration processes of the Republic of Macedonia, their continuous and persistent scientific monitoring and elaboration and active participation of MASA members in the preparation for the accession negotiations with the EU will remain a high priority on the agenda of MASA. Our ultimate goal is the Republic of Macedonia to become a democratic, economically prosperous and multicultural European country.
- The increasing incorporation of the international dimension in the scientific and artistic work of MASA, through the cooperation with foreign academies, scientific societies and other scientific institutions, through application and work on scientific projects financed by the European funds and the funds of other international financial institutions, also remains our important priority.

Let us congratulate ourselves on the great jubilee – 50 years of the Macedonian Academy of Sciences and Arts.

Прилози, Одд. за тех. науки, МАНУ, XXXV, 1–2 (2014), стр. 5–50
Contributions, Sec. Tech. Sci., MANU, XXXV, 1–2 (2014), pp. 5–50
ISSN 1857–6656
UDC: 551.509:[004.942:517.93

SHORT-TERM AND LONG-TERM FORECASTING WITH THE LORENZ 96 FAMILY OF MODELS

Igor Trpevski, Lasko Basnarkov, Borko Jovanovski, Zoran Utkovski, Ljupco Kocarev

*Research center for Computer Science and Information Technologies,
Macedonian Academy of Sciences and Arts*

Abstract: Contemporary numerical weather prediction schemes are based on ensemble forecasting. Ensemble members are obtained by taking different (perturbed) models started with different initial conditions. We introduce one type of improved model that represents interactive ensemble of individual models. The improved model's performance is tested with the Lorenz 96 model. One complex model is considered as reality, while its imperfect models are taken to be structurally simpler and with lower resolution. The improved model is defined as one with tendency that is weighted average of the tendencies of individual models. The weights are calculated from past observations by minimizing the average difference between the improved model's tendency and that of the reality. It is numerically verified that the improved model has better ability for short term prediction than any of the individual models. Furthermore we show how the imperfect models can be corrected by a simple parameter change of the dissipation coefficient so that a better long-term prediction is obtained.

1. Introduction

In forecasting the state of the atmosphere the occurrence of two fundamental types of errors is inevitable [1]. The first type, often called *internal error growth*, results from the amplification of initial condition uncertainties due to atmospheric instabilities [2]. The present tools for tackling internal error growth include data assimilation techniques and ensemble forecasting [3, 4, 5]. The second type is called external or model error and comes from the fact that atmosphere has larger complexity and resolution than its representations by models [6, 7]. With mathematical language, the atmosphere has (much) more degrees of freedom, or variables, and the equations of its evolution (if they exist) have more complex structure than those of the atmospheric models. Furthermore, sub grid-scale atmospheric processes such as cloud formation are not resolved by the models. Instead, they are approximated by different parametrization schemes, the choice of which mainly depends on the preference of the investigators at a weather forecasting center. The uncertainties coming from model errors limit our ability to make useful predictions with any individual model and current tools that account for model error are typically called empirical correction techniques [8, 9, 10, 11, 12].

Here we investigate the effects that unresolved processes and the lower resolution of the model have on both short-term (weather) and long-term (climatological) prediction with the Lorenz 96 family of models.

2. The Lorenz models

Lorenz [13] introduced a family of models representing an unspecified scalar atmospheric quantity at N equally spaced points about a latitude circle. Although artificial these models share some basic properties such as damping, advection and forcing and have proven very useful for studying and testing different techniques for data assimilation [14], and for exploring model errors in multi-scale systems [6]. There are three versions of the model with increasing complexity. The basic one – version I – simply captures the chaotic nature of the atmosphere and gives solution profile with irregular traveling waves. We will not use it and so its definition is skipped. Here we will use Lorenz model II as a prognostic model, while the Lorenz model III (which has multiple scales) will serve us as the ground truth.

The Lorenz model II expresses smooth propagating waves at N equally spaced points by representing the dynamics at each point using the following equation:

$$dX_n/dt = [X, X]_{K,n} - X_n + F, \tag{1}$$

where – X_n is the damping term and F represents forcing. The bracket operator repre- sents the double summation:

$$[X, Y]_{K,n} = \sum_{j=-J}^{J}{}' \sum_{i=-J}^{J}{}' (-X_{n-2K-i}Y_{n-K-j} + X_{n-K+j}Y_{n+K+j})/K^2, \tag{2}$$

where the symbol$\sum'$ denotes a modified summation, which is the same as the ordinary summation except that the first and last terms are divided by 2. One chooses a number K much smaller than N and we set $J = K/2$ if K is even and $J = (K - 1)/2$ if K is odd. The modified summation is used when K is even while the ordinary one is used when K is odd.

Version III of the Lorenz models superimposes fast, small-amplitude wave dynamics on top of the large-amplitude slow waves. The change in the dynamical variable that includes these two scales (for the n-th point on the latitude circle) is given by:

$$dZ_n/dt = [X, X]_{K,n} + b^2[Y, Y]_{1,n} + c[Y, X]_{1,n} - X_n - bY_n + F, \tag{3}$$

where b and c are coupling parameters. The slow and fast variables are obtained by using the equations:

$$X_n = \sum_{i=-I}^{I}{}' (\alpha - \beta|i|)Z_{n+i} \tag{4}$$

$$Y_n = Z_n - X_n. \tag{5}$$

The integer I and parameters α and β in the last equation are chosen so that the slow variable X_n is effectively a smoothed version of Z_n. The parameters suggested by Lorenz [13] are:

$$\alpha = (3I^2 + 3)/(2I^3 + 4I), \tag{6}$$

$$\beta = (2I^2 + 1)/(I^4 + 2I^2). \tag{7}$$

Note that for $\alpha = 1$, $\beta = 1$ and $I = 1$ model III reduces to model II.

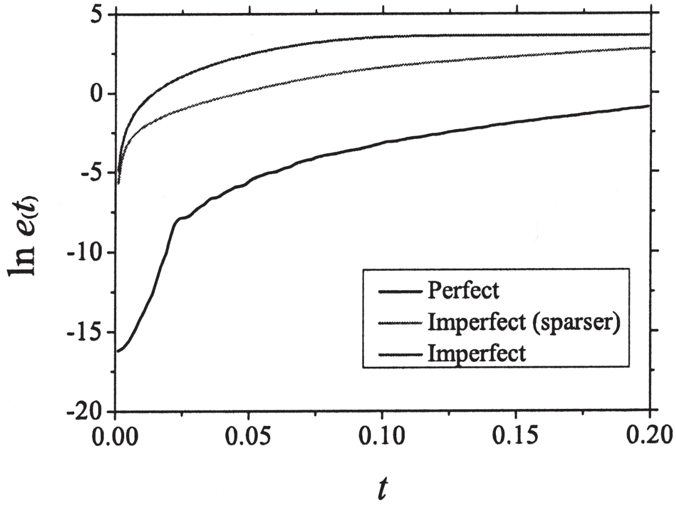
In the following numerical experiments we chose the ground truth to be given by (3) with $N = 960$ points, $K = 32$, $I = 12$, $b = 10$, $c = 2.5$. The value of the forcing will be determined by the experimental setup for the questions that we wish to answer. We will run (1) with different number of points to explore the effects of the model resolution on making predictions.

It is well established that the trajectories of nonlinear dynamical systems starting from very close initial conditions typically separate exponentially fast. But that happens if both trajectories are obtained from the same dynamical system. However, in reality the truth and the system belong to different classes of functions [6] and the divergence between them must not be exponential in the beginning. Comparison of the solutions of a model and the truth (and their mismatch) is the proper measure of the predictability power of the model. In figure 1 are shown root mean square difference between the truth and three different models of it - model with the same complexity M3, and two models of class II with different number of gridpoints. As can be seen only the difference is exponentially increasing only when the model is the same class of function as the truth. In the lower panel of the same figure are shown the same differences but in linear plot, where is clear that the growth of the error is linear [6].

3. Short-term forecast improvement of imperfect models

In this section we review a recently proposed method [12] which improves the short-term forecast of Model III by combining linearly the tendencies of several Model II systems with different parametrizations. The existence of tendency error suggests that possibly a linear combination of the tendencies of the models can give a tendency closer to the one of the truth. This is the idea behind the improvement of models. For every gridpoint of the models the improved model is a model with a tendency that is weighted combination of the tendencies of the individual models. More formally, if the tendency at gridpoint *n* of the model μ is

$$\dot{X}_n^\mu = T_n^\mu = [X^\mu, X^\mu]_{K,n} - X_n^\mu + F_n^\mu, \tag{8}$$



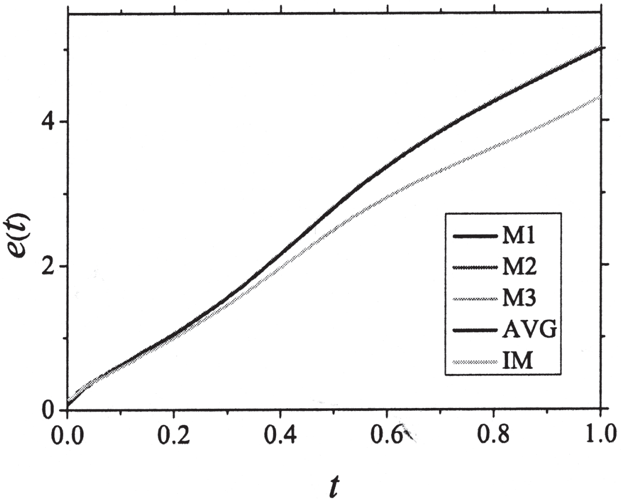


Figure 1: Growth of the error between the model and the truth. In the upper figure the ordinate axis has logarithmic scale and verifies the exponential divergence between the trajectories when the model is perfect (black curve). From the lower figure is clear that the error grows linearly - the improved model's curve is the lowest one then the improved model at gridpoint *n* has tendency

$$\dot{X}_n^s = T_n^s = \sum_\mu w_n^\mu T_n^\mu. \tag{9}$$

Assuming that the main source of limitation of the prediction is the error in deter- mination of tendency of the truth, one should use the tendency error as a measure of quality of a improved model. The average tendency error is given by

$$D = \langle \sum_{n=0}^{M-1} |T_n^t - T_n^s|^2 \rangle = \sum_{n=0}^{M-1} \langle \left| T_n^t - \sum_\mu w_n^\mu T_n^\mu \right|^2 \rangle, \tag{10}$$

where the tendency of the truth T^t

is given by the rhs of eq. (3) and angle brackets

denote time average. Optimal weights (according to the training set of data) are obtained by differentiating the last expression with respect to the weights

$$\begin{aligned} \frac{\partial D}{\partial w_n^\mu} &= \frac{\partial \langle |T_n^t - \sum_\nu w_n^\nu T_n^\nu|^2 \rangle}{\partial w_n^\mu} \\ &= 2\langle T_n^\mu \left(T_n^t - \sum_\nu w_n^\nu T_n^\nu \right) \rangle = 0. \end{aligned} \tag{11}$$

To simplify the notations one could introduce the covariances between the tendencies

$$\begin{aligned} C_n^{\mu,\nu} &= \langle T_n^\nu T_n^\mu \rangle, \\ C_n^{\mu,t} &= \langle T_n^t T_n^\mu \rangle. \end{aligned} \tag{12}$$

Then the equations for optimal weights at every gridpoint *n* become linear

$$\sum_{\nu} C_n^{\mu,\nu} w_n^{\nu} - C_n^{\mu,t} = 0, \tag{13}$$

where the factor 2 was removed with cancelation. The system of equations (13) can be written more succinctly with using matrix of covariances between the models $\mathbf{C}_n$, vector of covariances with the truth $\mathbf{c}_n$ and vector of weights $\mathbf{w}_n$ at every gridpoint n

$$\mathbf{C}_n \mathbf{w}_n = \mathbf{c}_n. \tag{14}$$

The linear regression technique suggests adding regularization term to avoid over- fitting of the parameters – weights in our case [15]. Then instead of minimizing only the average tendency error (10), the function to be minimized has the form

$$D + \lambda \sum_{n,\nu} (w_n^{\nu})^2, \tag{15}$$

where λ is the regularization coefficient. The minimization is obtained again by taking partial derivatives with respect to the weights. Then the system of equations for the weights (13) will have slightly modified form

$$\sum_{\nu} \left(C_n^{\mu,\nu} - \lambda\right) w_n^{\nu} - C_n^{\mu,t} = 0. \tag{16}$$

Using matrix notation, one concludes that for every gridpoint n the following matrix equation should be solved

$$\left(\mathbf{C}_n - \lambda \mathbf{I}\right) \mathbf{w}_n = \mathbf{c}_n, \tag{17}$$

which has a solution

$$\mathbf{w}_n = \left(\mathbf{C}_n - \lambda \mathbf{I}\right)^{-1} \mathbf{c}_n. \tag{18}$$

To perform numerical experiments on a PC we have considered as truth the Lorenz model III with $N = 960$ gridpoints. To have a more real setting we assume spatially dependent forcing F_n. In order to have a smoothly varying forcing we took perturbation of the constant $f_0 = 15$ that has randomly chosen Fourier components up to the order 10, while the higher were taken to be zero. More precisely the forcing is given by the sum

$$F_n = f_0 \left[1 + \sum_{m=1}^{10} f_m^c \cos\left(\frac{2\pi mn}{N}\right) + f_m^s \sin\left(\frac{2\pi mn}{N}\right) \right], \qquad (19)$$

where the spectral components ancf_m^c havf_m^sandom values from the interval $[-0.5, 0.5]$.

Also we assume that the models have different values of the forcing from the truth and between themselves as well. That can be represented if to the forcing of the truth is added another sum with random coefficients for each model.

Because any model of the atmosphere is its coarse representation we have taken M = 60 gridpoints for the models. For comparison of the models and the truth it was considered that the measurements are performed only at the gridpoints of the models. In calculations of the covariances we have assumed that the tendency of the truth is known. In reality the tendency of the atmosphere can be estimated with interpolation and the estimation will be different from the true value. To incorporate this fact we have added noise to the tendency of the truth. For short term forecasting purposes the models should be initiated from the state of the truth, and again with some perturbation that models the observation noise.

Within meteorological scientific community a measure for estimation of the pre- dictability range of a model is the anomaly correlation - AC [11]. AC for two time series is simply defined as a correlation between the two variables at the same moment. It measures how much on average, the deviation from their respective means at the same moment are at same direction and with similar magnitude. The AC between the truth and any model μ is given by

$$\mathrm{AC}^\mu = \frac{\sum_{m=0}^{M-1} \langle (Z_m - \langle Z_m \rangle)(X_m^\mu - \langle X_m^\mu \rangle) \rangle}{\sqrt{\sum_{m=0}^{M-1} (Z_m - \langle Z_m \rangle)^2} \sqrt{\sum_{m=0}^{M-1} (X_m^\mu - \langle X_m^\mu \rangle)^2}}. \qquad (20)$$

The angular brackets in the last equation again denote time averaging - in this case averaging is performed in the examination period. The predictability range extends to the moment when AC falls below value 0.6. In figure 2 we show the AC for the individual models, their average (calculated in the same way as for the prediction error) and the improved model. By using the threshold AC = 0.6 as a criterion for the predictability it is obtained that the improved model extends the predictability window for 17%. In

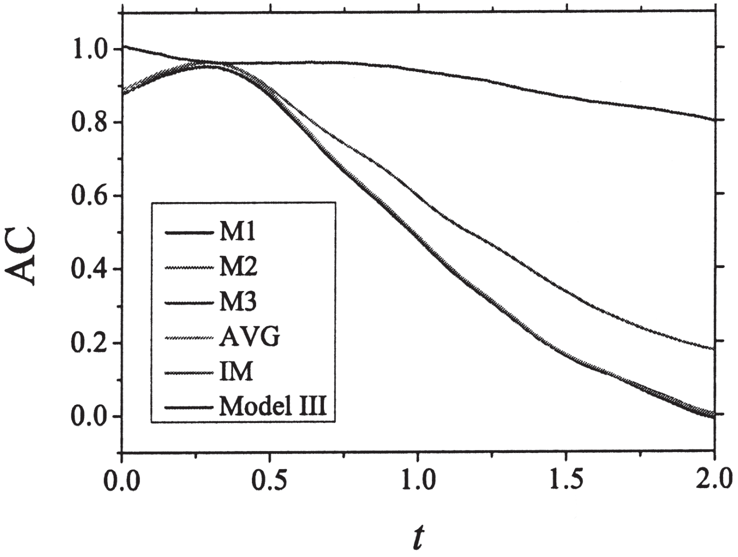


Figure 2: Anomaly correlation between the truth and the models. Top curve (in blue) is for model that has same complexity as the truth - model III. The middle curve (in green) corresponds to the improved model, and lower curves (almost indistinguishable) are for the individual models and average output of them.

The same figure is shown also the AC between the truth and a model that is the same as truth, started with close initial condition. That curve has typical behavior because it is decreasing. We think that the other AC curves first increase and then decrease due to the structural difference between the truth and the models.

4. Climatological prediction experiments

In this section we investigate the behaviour of the models in a climate change setting where the forcing is doubled over the course of the simulation. In this sense the ex- periments performed here are analogous to the transient climate sensitivity experiments performed with state-of-the-art coupled global circulation models. The essential differ- ence is the influence of the coupled ocean model in the latter, which has a lot of capacity to absorb heat during the course of one century.

We performed climatological experiments that explore the change in the global annual mean for the different models by gradually increasing their forcing over the course of 100 simulated years. The equation for the forcing is given by:

$$F(t) = F_{init} + \frac{t}{100 * MTU} 2F_{init} \tag{21}$$

where the term $100 * MTU$ denotes the number of model time units in the course of 100 years. The end value of the forcing is thus twice its initial value. Note that for the climatology experiments we didn't use spatial inhomogeneity in the forcing (19).

A single truth run was performed using model III with the same size and parameters as in the previous section. The initial value of the forcing is taken to be F_{init} = 15 for both the truth and the models. We used two versions of model II with different sizes, N = 240 and N = 480. One can expect different response to the increased forcing because of the different resolution. Also for each model version we run an ensemble of 40 members initialized from perturbed initial conditions obtained from the truth and calculate the global annual mean for each one. The results for these simulations are shown in blue in figures 3 and 4. Notice that regardless of the resolution of the model, the ensembles capture the global average of the truth run during the first two decades in each simulation. The same cannot be said for the remaining part of the simulations. Further increase in the forcing results in lower responses in the models than in the truth. Also the lower resolution of the model results in a smaller response to the increased forcing.

A plausible explanation to the lower response of model II is that it doesn't account for the small-amplitude fast process that is present in model III. In particular, for model III some of the energy is dissipated through the fast process. Since the fast process is unresolved in our model II we tested this hypothesis by increasing the dissipation coefficient. We repeated the simulations described previously but with the dissipation coefficient set to 1.2 and the results for these are shown in red in figures 3 and 4. We readily observe that the response to the increased forcing of both the $N = 240$ and $N = 480$ versions of model II is higher although the ensembles cannot fully capture the truth, especially for the later part of the century.

This indicates that one has to account for the unresolved processes by introducing a proper ’parametrization scheme’. Currently, the most promising techniques for this are the perturbed parameter approach [16] and stochastic paramterizations [17, 18]. The perturbed parameter approach takes uncertain parameters in the parametrization schemes and varies their values within their physical range. Alternatively, a stochastic scheme is one in which random numbers are included in the computational equations of motion so as to provide all possible realizations of the sub-gridscale motion.

5. Conclusion

We explored weather forecasting and climatological skill problems with the Lorenz 96 family of models. We have reviewed a recently proposed method for improving weather forecasting which intelligently combines the tendencies of several imperfect versions of the class II Lorenz 96 models. The improved model outperforms each of the imperfect models as well as their mean. One possible direction for further research is to attempt to apply these results in more real atmospheric models, or even for those that are used for numerical weather prediction. The main obstacle can be estimation of the tendency of the atmosphere. We think this kind of combination of state-of the-art models is worth testing because of importance of the weather prediction.

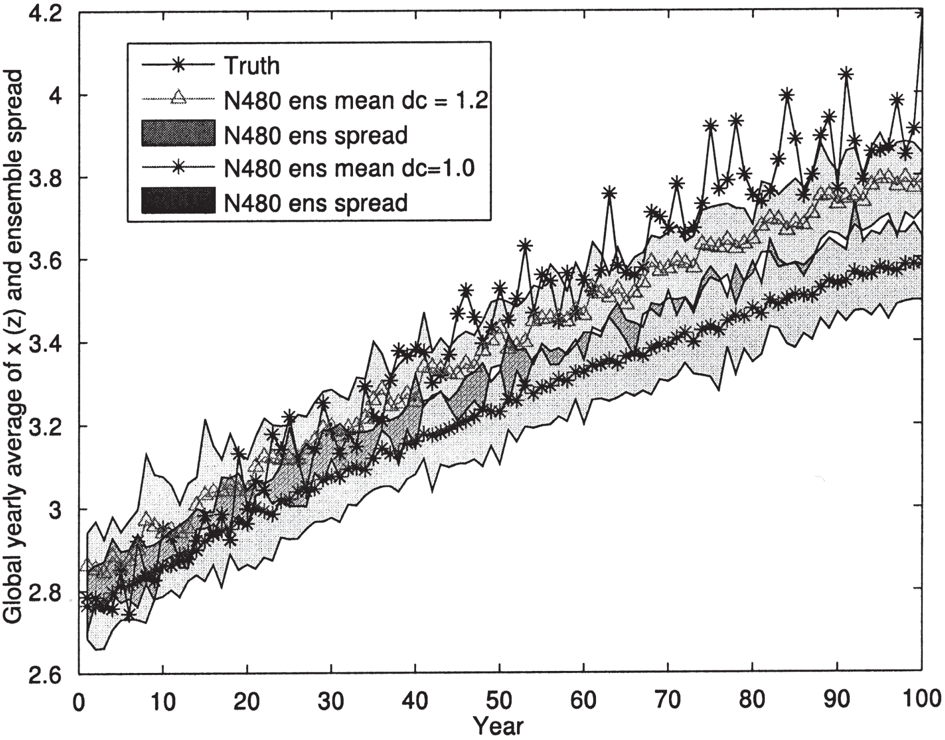


Figure 3: One hundred years simulation of model 2 with N=240 degrees of freedom with and without correction in the dissipation coefficient.

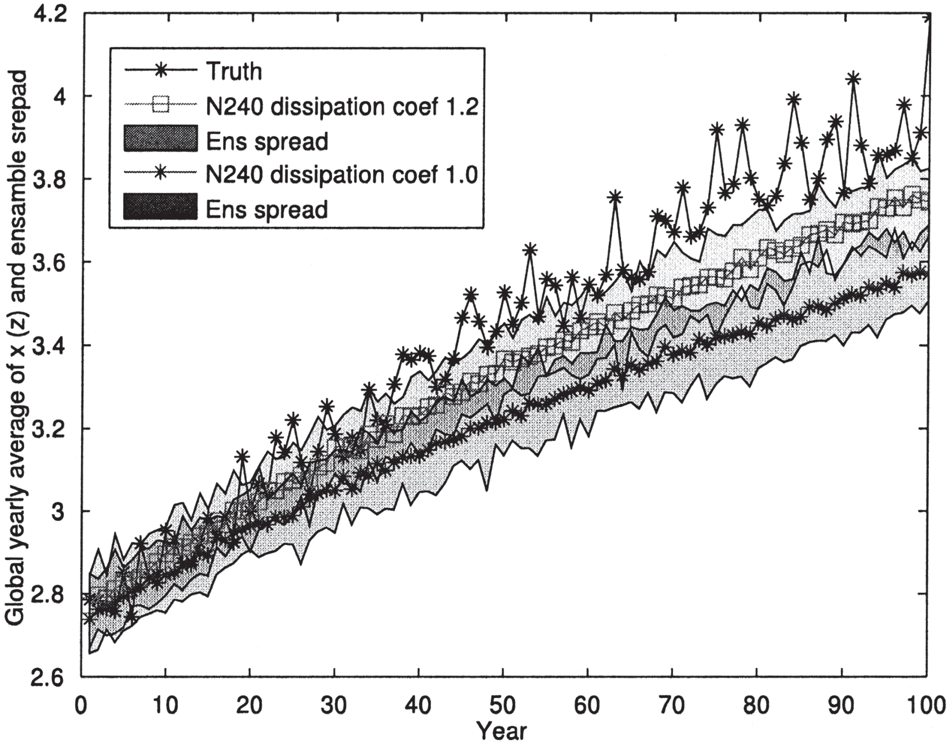


Figure 4: One hundred simulation of model 2 with N=480 degrees of freedom with and without correction in the dissipation coefficient.

Lastly, we analyzed the limitations of using imperfect models in climate change ex- periments. It was demonstrated that the dissipation of energy through the fast processes in the model 3 is in large measure responsible for the higher response to the increased forcing. Increasing the dissipation coefficient in the model 2 version resulted in a higher response. We indicate that instead of this ad-hoc change in the dissipation coefficient one should use a proper parametrization scheme. We argue that properly accounting for the unresolved processes in imperfect models present the most serious problem for climate change experiments.

References

[1] T. Palmer, “Predicting uncertainty in forecasts of weather and climate,” *Reports on Progress in Physics*, vol. 63, no. 2, p. 71, 2000.

[2] E. Kalnay, *Atmospheric modeling, data assimilation and predictability*. Cambridge university press, 2002.

[3] J. L. Anderson, “An ensemble adjustment kalman filter for data assimilation,” *Monthly Weather Review*, vol. 129, no. 12, pp. 2884–2903, 2001.

[4] B. Hunt, E. Kalnay, E. Kostelich, E. Ott, D. Patil, T. Sauer, I. Szunyogh, J. Yorke, and A. Zimin, “Four-dimensional ensemble kalman filtering,” *Tellus A*, vol. 56, no. 4, pp. 273–277, 2004.

[5] D. Merkova, I. Szunyogh, and E. Ott, “Strategies for coupling global and limited- area ensemble kalman filter assimilation,” *Nonlinear Processes in Geophysics*, vol. 18, no. 3, pp. 415–430, 2011.

[6] D. Orrell, L. Smith, J. Barkmeijer, and T. Palmer, “Model error in weather fore- casting,” *Nonlinear processes in geophysics*, vol. 8, no. 6, pp. 357–371, 2001.

[7] K. Judd and L. Smith, “Indistinguishable states ii: The imperfect model scenario,” *Physica D: nonlinear phenomena*, vol. 196, no. 3, pp. 224–242, 2004.

[8] C. Leith, “Objective methods for weather prediction,” *Annu Rev Fluid Mech*, vol. 10, no. 1, pp. 107–128, 1978.

[9] T. DelSole and A. Hou, “Empirical correction of a dynamical model. part i: Fun- damental issues,” *Mon Weather Rev*, vol. 127, no. 11, pp. 2533–2545, 1999.

[10] C. Danforth, E. Kalnay, and T. Miyoshi, “Estimating and correcting global weather model error,” *Mon Weather Rev*, vol. 135, no. 2, pp. 281–299, 2007.

[11] N. Allgaier, K. Harris, and C. Danforth, “Empirical correction of a toy climate model,” *Phys Rev E*, vol. 85, no. 2, p. 026201, 2012.

[12] L. Basnarkov and L. Kocarev, “Forecast improvement in lorenz 96 system,” *Nonlin. Processes Geophys*, vol. 19, pp. 569–575, 2012.

[13] E. N. Lorenz, “Designing chaotic models,” *Journal of the atmospheric sciences*, vol. 62, no. 5, pp. 1574–1587, 2005.

[14] Y.-n. Yoon, B. R. Hunt, E. Ott, and I. Szunyogh, “Simultaneous global and limited- area ensemble data assimilation using joint states,” *Tellus A*, vol. 64, 2012.

[15] C. Bishop, *Pattern recognition and machine learning*. Springer New York, 2006.

[16] J. Rougier, D. M. Sexton, J. M. Murphy, and D. Stainforth, “Analyzing the climate sensitivity of the hadsm3 climate model using ensembles from different but related experiments,” *Journal of Climate*, vol. 22, no. 13, pp. 3540–3557, 2009.

[17] D. S. Wilks, “Effects of stochastic parametrizations in the lorenz’96 system,” *Quar- terly Journal of the Royal Meteorological Society*, vol. 131, no. 606, pp. 389–407, 2005.

[18] H. Arnold, I. Moroz, and T. Palmer, “Stochastic parametrizations and model un- certainty in the lorenz’96 system,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1991, 2013.

Прилози, Одд. за тех. науки, МАНУ, XXXV, 1–2 (2014), стр. 5–50
Contributions, Sec. Tech. Sci., MANU, XXXV, 1–2 (2014), pp. 5–50
ISSN 1857–6656
UDC: 004.85:[004.63:811.163.3

OPINION MINING OF TEXT DOCUMENTS WRITTEN IN MACEDONIAN LANGUAGE

Andrej Gajduk* and Ljupco Kocarev*‡§

Keywords—opinion mining, classification, natural lan- guage processing, Macedonian language

Abstract: The ability to extract public opinion from web portals such as review sites, social networks and blogs will enable companies and individuals to form a view, an attitude and make decisions without having to do lengthy and costly researches and surveys. In this paper machine learning techniques are used for determining the polarity of forum posts on kajgana which are written in Macedonian language. The posts are classified as being positive, negative or neutral. We test different feature metrics and classifiers and provide detailed evaluation of their participation in improving the overall performance on a manually generated dataset. By achieving 92% accuracy, we show that the performance of systems for automated opinion mining is comparable to a human evaluator, thus making it a viable option for text data analysis. Finally, we present a f20ew statistics derived from the forum posts using the developed system.

* Macedonian Academy of Sciences and Arts.

‡ Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering, Skopje.

§ BioCircuits mstitute, University of California at San Diego.

1. Introduction

The World Wide Web (Web) has tremendously influenced our lives by changing the way we manage and share the information. Today, we are not only static observers and receivers of information, but in turn, we actively change the information content and/or generate new pieces of information. In this way, the entire com- munity becomes a writer, in addition to being a reader. Different mediums, such as blogs, wikis, forums and social networks, exist in which we can express ourselves by posting information and giving opinion on various subjects, ranging from politics and health to product reviews and travelling.

Sentiment analysis (also referred as opinion mining) concerns application of natural language processing, computational linguistics, and text analytics to identify and extract subjective information in source materials. Opinion Mining operates at the level of documents, that is, pieces of text of varying sizes and formats, e.g., web pages, blog posts, comments, or product reviews.

We assume that each document discusses at least one topic, that is, a named entity, event, or abstract concept that is mentioned in a document. Sentiment is the authors attitude, opinion, or emotion expressed on a topic. Although sentiments are expressed in natural language, they can in some cases be translated to a numerical or other scale, which facilitates further processing and analysis. Since the palette of human emotions is so vast and it is hard to select even the basic ones, most of the authors in the NLP community work with representation of sentiments according to their polarity, which means positive or negative evaluation of the meaning of the sentiment.

It is now well-documented that the opinions/views expressed on the web can be influential to readers in forming their opinions on some topic [1], and therefore, they are an important factor taken into consideration by product vendors [2] and policy makers [3]. There exists evidence that this process has significant economic effects [4]–[6]. Moreover, the opinions aggregated at a large scale may reflect political preferences [7], [8] and even improve stock market prediction

[9]. For the recent surveys on sentiment analysis or opinion mining we refer readers to [10]–[12].

The outline of the paper is as follows. In Section 2 the problem of opinion mining is formally defined. The proposed approach is outlined in Section 3. In Section 4 we give details about the datasets used in our experiments. In Section 5 the performance achieved using the different feature representation, classifiers and other text processing techniques is compared. A few statistics on the forum posts on kajgana derived using opinion mining are presented in Section 6. Section 7 concludes this paper.

2. Problem definition

In our experiment, we accept the classification of opin- ions according to their polarity i.e. polarity classification, used by the majority of authors [2], [10]. Pang and Lee [10] define polarity as the point on the evaluation scale that corresponds to our *positive* or *negative* evaluation of the meaning of the expressed opinion. However, not all texts are opinionated, so the method proposed by [13] which rates subjectivity and polarity separately is used. The problem is defined as follows:
Given a piece of text, decide whether it is subjective or objective, then assuming that the overall opinion in it is about one single issue or item, classify the opinion in subjective posts as falling under one of the two categories: positive or negative.

3. Proposed approach

A. Data representation

Text data in machine learning is commonly rep- resented by using the bag-of-features method [14]– [17]. This method is described as follows: let $D = \{f_1, . . . , f_m\}$ be a predefined set of m features that can appear in a forum post. We will refer to D as a feature dictionary. The features in the dictionary can be unigrams i.e. words such as *great* and *wasteful*, bigrams i.e. word pairs such as *not comfortable* or n-grams in the general case. Every post is represented by a vector of real numbers which correspond to a single feature in the feature dictionary. These values are computed using four different feature metrics.

- n-gram presence

$$presence_i^p = \begin{cases} 1, & \text{if } t_i^p \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

- η-gram count

$$count_i^p = t_i^p \tag{2}$$

- n-gram frequency

$$freq_i^p = \frac{t_i^p}{\sum_j t_j^p} \tag{3}$$

- n-gram frequency-inverse document frequency

$$ifreq_i^p = freq_i^p \log \frac{||P||}{||P_i||} \tag{4}$$

In (l-4) t_i^p is the number of occurrences of the ith η-gram in the post p, P is a set of all the posts and Pi is a set of posts containing at least one occurrence of the ith η-gram. Unigrams are the most commonly used in text mining, although some authors [18] recommend using bigrams. Their arguments include dealing with word negation and emphasizing which are very important in the domain of polarity classification.

B. Classifiers

The proposed feature metrics are evaluated using the two classifiers preferred by the majority of researchers in text classification [19] – [22].

- Support Vector Machines
- Naive Bayes

As discussed earlier, the classification will take place in two phases. First, subjectivity classification is performed where the comment is rated as either subjective or objective. Then if the post is subjective, it is classified as being either positive or negative. The latter will be referred to as polarity classification.

C. Preprocessing

Stop Words: Filtering stop words is a common practice in text mining [23]-[26]. Stop words are words with no informational value, such as function and lexical words. A suitable list of stop words in Macedonian language is difficult to obtain so one had to be manually prepared for this experiment. The list of stop words constitutes of 170 entries.

Stemming: Stemming has been extensively used to increase the performance of information retrieval systems for many international languages such as: English, French, Portuguese, to name a few [27], [28]. Stemming is a technique which aims to reduce a word to its stem or root form. Thus, literally different words that share a common stem may be abstracted as a single informational entity. There are several common approaches to stemming as categorized in [29] namely affix removal method, successor variety method, n-gram method and table lookup method. Affix removal which includes algorithms such as Lovins or Porter, is the most popular method, but relies heavily on manually defined rule sets. A good rule set for Macedonian is yet to be defined, which is why we decided to use a stemming method that relies on nothing more but the set of words that need to be stemmed. This method is called peak-and-plateau and is based on tries. For a more detailed explanation to this method we refer readers to [30].

D. Rule bigrams

Some authors propose a different way of incorporating bigrams into the feature vector [31]–[34] which will be refereed to as *rule bigrams*. According to this approach all negatory words are appended a tag *e.g. not* to the word following the negatory word in the sentence. Thus the bigram *not good* becomes the unigram *notgood.* This method is adopted and expanded to emphasizery words, thus transforming bigrams such as *most disgusting* and *very disgusting* into the same unigram *e.g. verydisgust-ing.* This approach is adequate when using unigram presence as a feature vector, but we propose an alteration when applying it in combination with other feature met-rics that rely on counting the unigram occurrences. Any occurrence of an unigram preceded by an emphasizing word is counted as two occurrences of the corresponding unigram i.e. $\hat{t}_i^p = 2t_i^p$, whereas any occurrence of an unigram preceded by a negatory word is considered as -1 occurrence of the corresponding unigram i.e. $\hat{t}_i^p = -t_i^p$.

Accuracy	SVM	NB
Presence	0.76	0.64
Count	0.73	0.55
Frequency	0.72	0.61
IFrequency	0.94	0.78

TABLE I: Accuracy, no preprocessing

4. Dataset

The domain used in this study is forum posts which are written in Macedonian language from the kajgana forum. Forum posts tend to be less focused and organized than other text documents such as product reviews for instance, and consist predominantly of informal text. The posts on kajgana are grouped into 47 disjoint topics which are then divided into subtopics (over 50,000) and are 60 words long on average. There are a total of 4 million unique words in the posts. In our experiment, we ignored words that have less than 5 occurrences in order to reduce the total dictionary size and to eliminate type errors. This left us with 800,000 unique words. A total of 800 posts were manually tagged of which 260 are positive, 260 are negative and 280 are objective posts. This dataset will be used for evaluations on the different classifiers and feature representations. All evaluations are done using 10-fold cross validation to avoid over-fitting.

5. Results

First, the aforementioned feature representations using unigrams in combination with the two proposed classifiers are evaluated. Inverse frequency the best feature representation followed by presence (Table. I). As for classifiers, SVM outperforms NB on every feature representation.

Surprisingly, stemming and stop words removal reduces accuracy (Table II). More specifically the accuracy drops from 0.94 to 0.74 when using an SVM classifier

Accuracy	SVM	NB
Presence	0.76	0.63
Count	0.72	0.56
Frequency	0.70	0.60
IFrequency	0.74	0.62

TABLE II: Accuracy, with preprocessing

Presence	SVM	NB
Unigrams only	0.76	0.63
Bigrams only	0.54	0.52
Unigrams bigrams	0.79	0.67

IFrequency	SVM	NB
Unigrams only	0.74	0.62
Bigrams only	0.55	0.52
Unigrams bigrams	0.75	0.62

TABLE III: Accuracy, bigrams

and from 0.78 to 0.62 when using an NB classifier. One possible reason is that the word stemming algorithms does not perform well for the Macedonian language.

As mentioned earlier the proposed feature representations can be applied to η-grams of any size, although so far only unigrams have been used. Next, we evaluate presence and ifrequency using bigrams, alone and in combination with unigrams (Table III). Bigrams alone are not good features, but when used in conjunction with unigrams they show a slight improvement when presence as feature representation is used from 0.76 to 0.78 with SVM and from 0.63 to 0.67 with NB.

Finally, in Table IV the accuracy when using rule bigrams (only negation rules, only emphasis rules and both together) are given. The results show that rule bigrams do not impact classification accuracy, with the exception of negation rules that achieves a slight increase in accuracy for unigram presence.

6. Statistics

Using the combination of unigram ifrequency for a feature representation and SVM as a classifier some interesting properties of forum posts in general can be demonstrated. As stated above the forum posts are divided into several topics. Let us denote with *pt* the number of positive posts and with n_t the number of negative posts for each topic *t*. The overall mood on the topic m_t is defined as

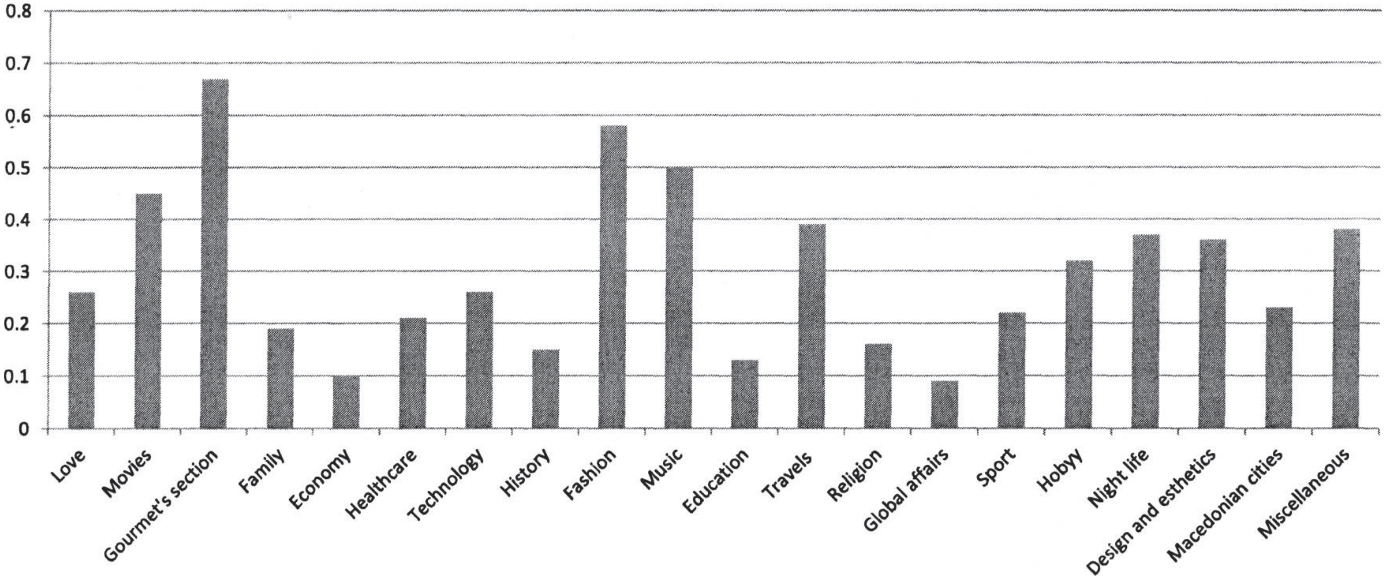


Fig. 1: Mood by topic

Presence	SVM	NB
Unigram	0.76	0.63
Negations only	0.78	0.62
Emphasizers only	0.76	0.61
Both	0.77	0.62

IFrequency	SVM	NB
Unigram	0.74	0.62
Negations only	0.73	0.59
Emphasizers only	0.74	0.59

TABLE IV: Accuracy, rule bigrams

Interestingly, people are most positive when discussing food (Gourmets section) and fashion, but are extremely negative on global affairs and the economy (Fig. 1). .

In a similar fashion the posts can be grouped and their mood calculated by month as displayed in Fig. 2. The public mood is highest in spring (May and April), probably due to the good weather during these two months.

7. Conclusion

In this paper forum posts written in Macedonian lan- guage are labeled as being positive, negative or objective. We show that this can be done with great accuracy using simple text feature extraction metrics such as unigram presence and standard classifiers such as Naive Bayes. The best accuracy is achieved by using a combination of unigram frequency-inverse document frequency for a feature metrics and support vector machines as a classi- fier: 0.96 on subjectivity classification, 0.96 on polarity

classification or a total classification accuracy of 0.92. Additionally, we tested various techniques for improving the performance. Of these, word stemming and stop words removal had a negative effect on classification accuracy. The use of bigrams does not help with the classification task while using rule bigrams increases the accuracy only slightly in polarity classification.

References

[1] Y. Lin, J. Zhang, X. Wang, and A. Zhou, “Sentiment classification via integrating multiple feature presentations,” in *Proceedings of the 21st international conference companion on World Wide Web*, pp. 569–570, ACM, 2012.

[2] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up?: sentiment classification using machine learning techniques,” in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pp. 79–86, Association for Computational Linguistics, 2002.

[3] J. A. Horrigan, “Online shopping,” *Pew Internet & American Life Project Report*, vol. 36, 2008.

[4] W. Antweiler and M. Z. Frank, “Is all that talk just noise? the information content of internet stock message boards,” *The Journal of Finance*, vol. 59, no. 3, pp. 1259–1294, 2004.

[5] N. Archak, A. Ghose, and P. G. Ipeirotis, “Show me the money!: deriving the pricing power of product features by mining consumer reviews,” in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 56–65, ACM, 2007.

[6] J. A. Chevalier and D. Mayzlin, “The effect of word of mouth on sales: Online book reviews,” tech. rep., National Bureau of Economic Research, 2003.

[7] T. Mullen and R. Malouf, “A preliminary investigation into sentiment analysis of informal political discourse.,” in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pp. 159–162, 2006.

[8] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, “Predicting elections with twitter: What 140 characters reveal about political sentiment.,” *ICWSM*, vol. 10, pp. 178–185, 2010.

[9] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[10] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[11] H. Tang, S. Tan, and X. Cheng, “A survey on sentiment detection of reviews,” *Expert Systems with Applications*, vol. 36, no. 7, pp. 10760–10773, 2009.

[12] M. Tsytsarau and T. Palpanas, “Survey on mining subjective data on the web,” *Data Mining and Knowledge Discovery*, vol. 24, no. 3, pp. 478–514, 2012.

[13] N. Godbole, M. Srinivasaiah, and S. Skiena, “Large-scale sentiment analysis for news and blogs.,” *ICWSM*, vol. 7, 2007.

[14] T. Nakagawa, K. Inui, and S. Kurohashi, “Dependency tree-based sentiment classification using crfs with hidden variables,” in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 786–794, Association for Computational Linguistics, 2010.

[15] J. Read, “Using emoticons to reduce dependency in machine learning techniques for sentiment classification,” in *Proceedings of the ACL Student Research Workshop*, pp. 43–48, Association for Computational Linguistics, 2005.

[16] M. J. Paul, C. Zhai, and R. Girju, “Summarizing contrastive viewpoints in opinionated text,” in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 66–76, Association for Computational Linguistics, 2010.

[17] M. Rushdi Saleh, M. T. Mart´ın-Valdivia, A. Montejo-Ra´ez, and
L. Uren˜a-Lo´pez, “Experiments with svm to classify opinions in different domains,” *Expert Systems with Applications*, vol. 38, no. 12, pp. 14799–14804, 2011.

[18] M. Zeng, Y. Yang, and W. Liu, “An approach of text sentiment analysis for public opinion monitoring system,” in *Semantic Web and Web Science*, pp. 131–141, Springer, 2013.

[19] T. Mullen and N. Collier, "Sentiment analysis using support vector machines with diverse information sources.," in *EMNLP*, vol. 4, pp. 412–418, 2004.

[20] Q. Ye, B. Lin, and Y.-J. Li, "Sentiment classification for chinese reviews: A comparison between svm and semantic approaches," in *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*, vol. 4, pp. 2341–2346, IEEE, 2005.

[21] M. Gamon, "Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis," in *Proceedings of the 20th international conference on Computational Linguistics*, p. 841, Association for Computational Linguistics, 2004.

[22] M. Koppel and J. Schler, "The importance of neutral examples for learning sentiment," *Computational Intelligence*, vol. 22, no. 2, pp. 100–109, 2006.

[23] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language resources and evaluation*, vol. 39, no. 2-3, pp. 165–210, 2005.

[24] P. Chaovalit and L. Zhou, "Movie review mining: A comparison between supervised and unsupervised classification approaches," in *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, pp. 112c–112c, IEEE, 2005.

[25] C. Strapparava and R. Mihalcea, "Semeval-2007 task 14: Affective text," in *Proceedings of the 4th International Workshop on Semantic Evaluations*, pp. 70–74, Association for Computational Linguistics, 2007.

[26] A. Esuli and F. Sebastiani, "Determining term subjectivity and term orientation for opinion mining.," in *EACL*, vol. 6, p. 2006, 2006.

[27] W. B. Frakes and C. J. Fox, "Strength and similarity of affix removal stemming algorithms," in *ACM SIGIR Forum*, vol. 37, pp. 26–30, ACM, 2003.

[28] J. Savoy, “Searching strategies for the hungarian language,” *Information processing & management*, vol. 44, no. 1, pp. 310–324, 2008.

[29] D. Sharma, “Stemming algorithms: A comparative study and their analysis,” *International Journal of Applied Information Systems*, vol. 4, no. 3, pp. 7–12, 2012.

[30] M. A. Hafer and S. F. Weiss, “Word segmentation by letter successor varieties,” *Information storage and retrieval*, vol. 10, no. 11, pp. 371–385, 1974.

[31] M. Ghiassi, J. Skinner, and D. Zimbra, “Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network,” *Expert Systems with Applications*, 2013.

[32] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, “Recursive deep models for semantic compositionality over a sentiment treebank,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013.

[33] S. Wang and C. D. Manning, “Baselines and bigrams: Simple, good sentiment and topic classification,” in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pp. 90–94, Association for Computational Linguistics, 2012.

[34] H. Kang, S. J. Yoo, and D. Han, “Senti-lexicon and improved na¨ıve bayes algorithms for sentiment analysis of restaurant reviews,” *Expert Systems with Applications*, vol. 39, no. 5, pp. 6000–6010, 2012.

Содржина – Contents

ON THE EVE OF THE GREAT JUBILEE – 50 YEARS OF THE MACEDONIAN ACADEMY OF SCIENCES AND ARTS 1967 – 2017 V
Taki Fiti, president of the Macedonian Academy of Sciences and Arts

SHORT-TERM AND LONG-TERM FORECASTING WITH THE LORENZ 96 FAMILY OF MODELS 5
Igor Trpevski, Lasko Basnarkov, Borko Jovanovski, Zoran Utkovski, Ljupco Kocarev

1. Introduction. 6
2. The Lorenz models 6
3. Short-term forecast improvement of imperfect models. 8
4. Climatological prediction experiments 14
5. Conclusion 15

References 18

OPINION MINING OF TEXT DOCUMENTS WRITTEN IN MACEDONIAN LANGUAGE. 21
Andrej Gajduk, Ljupco Kocarev

1. Introduction. 22
2. Problem definition 23
3. Proposed approach 23
4. Dataset 26
5. Results. 27
6. Statistics 28
7. Conclusion 31

References 32

Издавач:
Македонска академија на науките и уметностите – Скопје

Графичка и компјутерска обработка:
Давид Георгиевски

Печат:
„Бомат графикс“ – Скопје

Тираж:
300 примероци

Бројот е излезен од печат во март 2017 година.