Usefulness of Machine Learning technology for classification of research funding submissions

Nenad Cuturic, 700508-7874, DV2594 H20 lp12 Machine Learning for Streaming Data, Blekinge Institute of Technology

***Abstract*—This paper presents the result of an experimental study, which is a part of the university course, that investigates whether machine learning (ML) technology can be useful for the text topic classification of submissions' abstracts for research funding classified according to United Nation's 17 Global Goals.**

Labeled data is supplied by The Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (Formas) and Sweden's innovation agency (Vinnova) while complete though unlabeled data set can be downloaded from the Swedish national database (SweCRIS). Only texts in Swedish language were analyzed.

The result is showing that machine learning could be useful for the classification of the submissions but it is necessary to collect mach more reliable labeled data.

***Index Terms* —Natural Language Processing (NLP), Multinomial Naïve Bayes, topic mining, document classification**

I. Introduction

On the behalf of the Swedish Government the Swedish Research Council (VR) is managing SweCRIS which contains public information about how twelve participating research funding bodies have distributed their money to the Swedish recipients. As a part of the 2030 Agenda for Sustainable Development adopted by all UN's member states in 2015 VR is interested in investigating whether machine learning technology could be successfully used for estimation of how well submission in SweCRIS are meeting UN's 17 Global Goals.

The data in SweCRIS database is untagged. Two of the participating government agencies, Formas and Vinnova, have provided labels for data and this project aims to use those labels with content from SweCRIS for performing this study. Labels produced by Vinnova are based on self-classification by the applicants while the labels from Formas consist of self-classified data which is later on manually analyzed and re-classified resulting in the detailed report [1].

The report is showing that self-assessed labels are of low quality (unreliable) which makes this study even more important.

Quality level of the data content (submissions) is not analyzed.

The plan has been to use two ML techniques: Multinomial Naïve Bayes classification (MNB) and Support Vector Machines (SVM), and compare the results. Prior to ML some Natural Language Processing (NLP) of the texts needed to be done.

II. Natural Language processing and Machine learning

A. *Problem Definition*

As mentioned in the introduction the available data set is divided in the following groups:

1. Untagged (huge majority of the data set),
2. Tagged by labels with high quality (796), and
3. Tagged by labels with questionable quality (15.234).

Formas' evaluation of the self-assessed classification is showing that such type of labeling is unreliable. That is the reason why this (the third) type of labeling (from Vinnova) is used only when necessary to fill in the second type of the labeled data for the classes where there was insufficient number of labels to balance out the number of labels in each class. This is illustrated in Fig.1

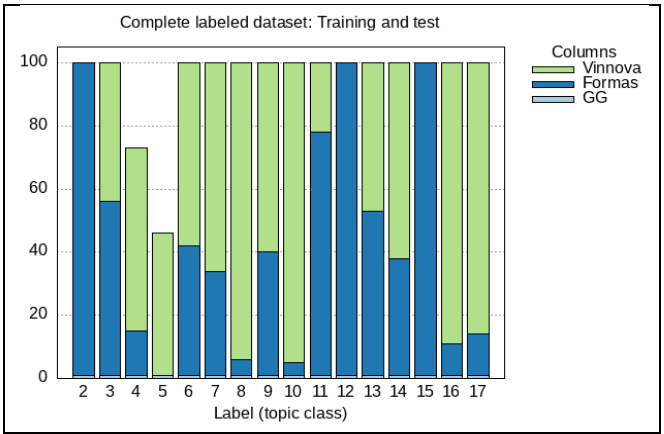

Fig. 1. Distribution of used labeled data set per class [2-17], at most 100 and source: Global Goals, Formas and Vinnova. Total set: 1534

GG in the figure stands for Globala Goals and its data comes from official description of goals scraped from https://www.globalamalen.se/ and for each goal there is one document in data set.

Due to insufficient number of labeled documents for class 1, only 8 from all sources, that class is excluded from the study. The number of documents per class is chosen to be at most 100 but despite that classes 4 and 5 didn't have sufficient number of labeled data. The numbers were greater than 50% so the decision is made to include them the study anyhow.

Labeled data set needs to be reasonably balanced per class for ML algorithms to be able to learn and predict result accurately. 100 is chosen so we could cover as many classes as possible and nevertheless have reasonable balanced data set distribution. Unbalanced data is leading to skewed results.

Unbalanced (and of low quality) labels have for example led to very low prediction accuracy in results for the class 5 which will be shown later in the text.

Another problem is that the same document can belong to more than one class. Indeed self-classified documents are labeled with the main (primary) and up to 2 secondary labels. Even the main classification has been shown to be unreliable so the choice has been made to use primary classification only.

The field of Natural Language Processing (NLP) is vast and there are a lot of different problem areas where different techniques are used for solving problems within different fields. The classification problem belongs to the "topic mining" field: Given the number of text documents and given the predetermined number of topics, analyze given texts and find out the probability that the specific text belongs to certain topic.

The highest probability with the possible threshold can then be used as a decider for decision about which topic analyzed text belongs to. The document can belong to none of the topics which threshold can help with or as alternative if we had sufficient number of labeled documents from class 0 (not belonging to any topic) we would be able to treat them as an ordinary (additional) class.

Prior to applying ML techniques the text data has to be pre-processed. In this study we are going to simplify things and assume that each document must belong to some of the identified topics.

The other simplification is that we are going to use documents as so called "bag of words" where each word in the text will be treated as an separate entity without the relationship to any other word in the sentence. This process is call word tokenization.

For ML algorithms to be applied we need to convert text to some kind of corresponding numerical representation that computer can work with. In other words we are only going to use statistical/probabilistic properties of the word distribution: frequency of the each word related to the word corpora and number of texts, and use some scoring technique like tf-idf (term-frequency inverse document frequency). Tf-idf technique is explicitly used in the processing pipeline as a part of the used ML-models.

As part of pre-processing and cleaning up we are also going to apply stop-words, another common technique for NLP, where some very common words like "the" in English - that are bringing no information but produce "noise" to the model - are removed from the text. Besides general stop-words, some of the identified stop-words specific to this corpora are added to the list.

Even other types of cleaning up of the text is done, like removing the interpunctions and stemming. Stemming is the process of reducing words to their word stem, base or root form, allowing us to group and analyze similar words as a single item.

Statistical representation (term distribution) of our word corpora after cleaning up is presented on Fig.2.

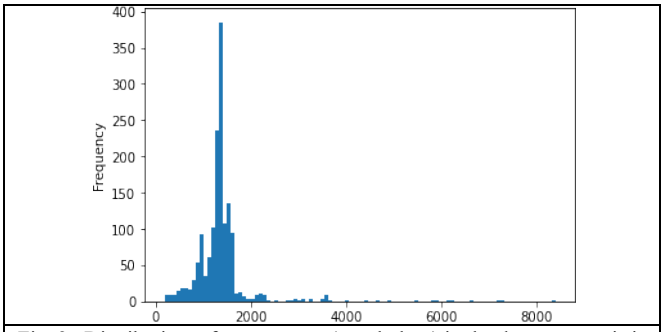

Fig. 2. Distribution of text corpora (vocabulary) in the data set, x-axis is corpora (unique words) while y-axis is number of appearances (frequency)

For machine learning part we are going to use the following two models:

- Multinomial Naïve Bayes (MNB)
- Support Vector Machines (SVM)

and compare the results.

B. Multinomial Naïve Bayes (MNB)

MNB is a specialized version of naïve bayes designed to handle text documents using word counts as its underlying method of calculating probability. It is chosen because of its simplicity. Combined with tf-idf algorithm it is powerful tool for solving topic mining problem.

C. Support Vector Machine (SVM)

SVM is another common supervised (labeled data) ML method for solving classification problems.
The objective of the SVM algorithm is to find a hyperplane in an N-dimensional space (N is the number of classes) that distinctly classifies the data points.
To separate the two classes of points, there is an infinitive number of hyperplanes that can be chosen. Our objective is to find "the" plane that has the maximum margin, i.e the maximum distance between points belonging to the different classes. Larger the margin distance larger the confidence which we can classify future documents with.
This method is also chosen because of its simplicity but strong capability to solve classification problems.
After analyzing and processing, the data results of those two different techniques are presented and compared.

III. Results

Common method for presenting the test results of the classification problem is using s.k. confusion matrix. Such a matrix has one dimension presenting labels for the test data which is the "truth" = prior, and the second dimension represents the predicted labels (posterior) by the trained model.
In the Fig.3. and 4. visual presentations of the confusion matrix for the MNB and SVM is shown in form of Heat map. In the monochromatic heat map intensity of the color represents the frequencies: more color corresponds to higher frequency.
Ideal model with all correct result would have only diagonal (i,i), i=1..N (N-number of classes) with values larger then 0. The more values outside the diagonal - the worse model is performing.

A clear visual illustration of the bad performing model is in the last raw of the Fig.3 for class 17.
The columns of the matrix for (i,j) where i ≠ j, i.e. outside the (i,i) diagonal, represent number of wrongly predicted classes for a given class.
This kind of method of presenting the results is more of the quantitative nature.
Looking at the figures we can clearly identify the distinct diagonal, which is a good indicator for estimating usefulness.

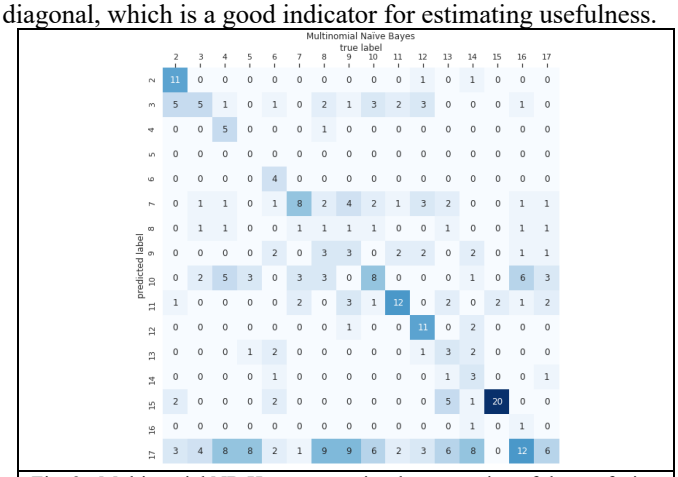

Fig. 3. Multinomial NB Heat map – visual presentation of the confusion matrix for the classification, diagonal represents correct predictions.

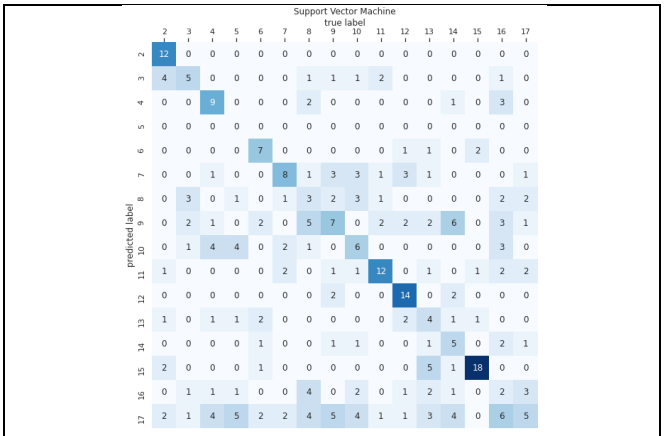

Fig. 4. SVM Heat map – visual presentation of the confusion matrix for the classification, the values on the diagonal represents correct predictions.

The other more qualitative way of presenting of results is by giving the following metrics:

$$\boldsymbol{precision} = \frac{true\ positives}{true\ positives + false\ postivives}$$

$$\boldsymbol{recall} = \frac{true\ positives}{true\ positives + false\ negatives}$$

$$\boldsymbol{F1-score} = 2 * \frac{precision * recall}{precision + recall}$$

$$\boldsymbol{support} = number\ of\ samples\ per\ class$$

$$\boldsymbol{accuracy}(\hat{y}_i, y_i) = \frac{1}{n_{samples}} \sum_{n=1}^{n_{samples}} 1(\hat{y}_i = y_i)$$

where

- $\hat{y}_i$ is the predicted value of the n-th sample and y_i is the corresponding true value.
- 1(x) is indicator function (1 if equal otherwise 0)

Metrics descriptions:

- Accuracy shows how good a model is at predicting the correct classes. If dataset is fairly balanced and every category has equal importance, this is the go-to metric to measure model's performance.
- The F1-score is weighted average of the precision and recall (1 is best).
- The precision is intuitively the ability of the classifier to avoid predicting negative value as positive.
- The recall is intuitively the ability of the classifier to find all the positive samples.

Fig.4. and 5. are the histogram representations of these metrics, except for accuracy, for MNB and SVM respectively.
Next table presents summary of all average values per class (shaded rows are from SVM-model).

avg	precision	recall	F1-score	accuracy	support
macro	0,44	0,33	0,32	0,33*	307
	0,41	0,37	0,37	0,38*	307
weighted	0,46	0,33	0,33	0,33*	307
	0,42	0,38	0,39	0,38*	307

* accuracy is a single value (has no average)

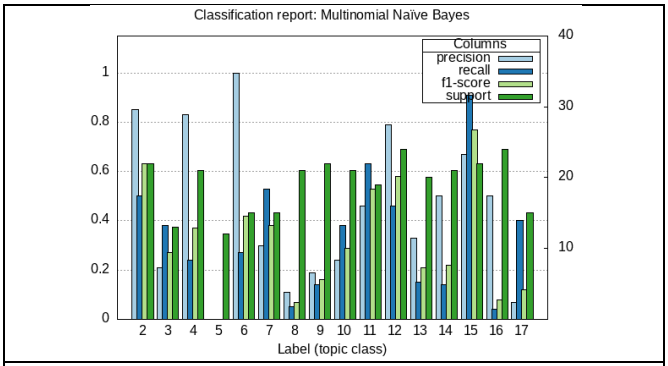

Fig. 5. Classification report: Multinomial Naïve Bayes, **support** is using the right y-axis (frequency count), while the rest are using the left one (%).

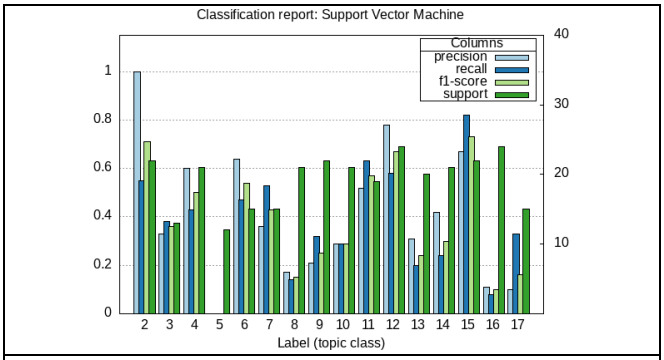

Fig. 5. Classification report: Support Vector Machines, **support** is using the right y-axis (frequency count), while the rest are using the left one (%).

IV. Why is quality of labels so important?

Looking at the results both models are showing signs of good tendency in predicting correct results, taking in consideration the low number of correctly labeled data.

Where the number of the quality-certified samples reaches the max number (100) predicting accuracy tends to perform very well (se class 2) while with the low quality of data and/or the low number of samples, models are performing purely (see class 5 and 17). Class 5 could basically be removed from the study due to its underperformance.
Class 4, that also has lower number of samples, is performing better then class 8, that has full set of samples but of lower quality than those in class 4. So it is more important to have quality (labels that correctly classify the data) then quantity (larger number of labeled documents though with questionable quality).
This is consistent with the conclusions Anders made in his report[1].

V. Which model performs better and wat metrics should be used?

Du to time limit, study hasn't investigate which metrics' are most important. Most of the metrics show the same tendency anyhow.

It is pretty safe to use accuracy as a single measure for comparison of how well models are performing.

Looking at all metrics values we can conclude that SVM-model performs slightly better. This could also be observed throughout several runs where SVM accuracy usually shows values that are 2 to 5 percentage points larger than those for MNB.

Both models are very quick on a computer with gpu-card supporting cuda so processing speed should not influence the model choice.

Based on all the data we can state that SVM-model is preferable.

VI. Conclusion

As a future improvement more correctly labeled data should be collected.

Input and intermediate data in itself could be analyzed by the expert(s) to improve list of stop-words), whether stemming is bringing more good than harm and how interpunctions are used. For instance it is observed that text is copy paced from other source into the form causing words in the end of the line broken in two parts. Cleaning of interpunctions and whitespaces is breaking the word in two parts which could easily cause loss of information. This seems to occur pretty based on few analyzed samples.

Du to time constraints only two models are used for evaluation but it would be really interesting to compare it to at least one more for instance neural networks.

It is intriguing to test some of the unsupervised models i.e. those treating documents as unlabeled data and with some clustering techniques trying to mine its own topics.

References

[1] Anders Clarhäll, *Formasfinansierade projekt och de globala hållbarhetsmålen* PM, Dnr: 2020-03160, Formas Org.nr. 202100-5232, Drottninggatan 89 Stockholm.