IRSTI 28.23.35

A. Serek[1], M. Zhaparov[2]
[1]Suleyman Demirel University, Kaskelen, Kazakhstan
[2]Paragon International University, Cambodia

SENTIMENT ANALYSIS OF UNIVERSITY FEEDBACK OPINION OF STUDENTS ABOUT AN EDUCATIONAL PART IN KAZAKH LANGUAGE USING MULTIBINOMIAL NAIVE BAYES CLASSIFIER

Abstract. In this paper, the system which identifies the sentiment, (aka meaning) of a kazakh phrase, (whether it is a positive, or a negative) have been implemented using MultiBinomial Naive Bayes Classifier and achieved accuracy approximately 71 % on the dataset about university feedback across students on its educational component in order to help administrative staff to evaluate the current state of education in the university and make some decisions on its basis. We consider it to be a good result, given that the data was small in size, so that there were only few collected samples. The importance of the work that we did not find any paper which performed sentiment analysis using MultiBinomial Naive Bayes classifier on an agglutinative language. It can be argued, that the model can be successfully generalized in other educational organizations pursuing the same cause as it was identified in the above-mentioned rationale. The limitation of the paper is that only one algorithm has been applied to it, and the dataset size is small.

Keywords: sentiment analysis on kazakh language, sentiment analysis, nlp on kazakh text, agglutinative languages, bayes classifier nlp.

***

Аннотация. В этой статье система, которая идентифицирует значение казахской фразы, (будь то положительный или отрицательный) были реализованы с использованием Мультибиномиального Байесовского классификатора и была достигнута точность в значении примерно 71% в наборе данных об отзывах студентов об образовательном компоненте университета, чтобы помочь административному персоналу оценить текущее состояние образования в университете и принять решения на его основе. Мы считаем, что это хороший результат, учитывая, что данные были небольшими по размеру, так что было собрано всего несколько образцов. Важность работы заключается в том, что мы не нашли ни одной статьи, в которой проводился анализ

настроений с использованием этого классификатора на агглютинативном языке. Можно утверждать, что модель может быть успешно обобщена в других образовательных организациях, преследующих ту же причину, которая была определена в вышеупомянутом обосновании. Ограничение статьи состоит в том, что к ней был применен только один алгоритм, а размер набора данных невелик.

Ключевые слова: анализ настроений на казахском языке, анализ настроений, nlp в казахском тексте, агглютинативные языки, байесовский классификатор nlp.

***

Аңдатпа. Бұл мақалада қазақ сөз тіркесінің мағынасын анықтайтын жүйе, (оң немесе теріс пе) Мультибиномиальды Байес классификаторын қолдану арқылы жүзеге асырылды және әкімшілік қызметкерлерге университеттегі ағымдық жағдайды бағалауға және соған негізделген шешімдер қабылдауға көмектесу үшін университеттің білім беру компоненті туралы студенттердің пікірлер жиынтығында шамамен 71% шамасында дәлдікке қол жеткізілді. Деректер аз болғандықтан, бұл жақсы нәтиже деп санаймыз, сондықтан бірнеше үлгілерді жинадық. Жұмыстың маңыздылығы агглютинативті тілде осы классификаторды қолдана отырып, көңіл-күйге талдау жасалынған бірде-бір мақаланы таба алмағанымызда. Үлгіні басқа білім беру ұйымдарында жоғарыда келтірілген негіздемеде анықталған себепке сүйене отырып ойдағыдай жалпылауға болады деп айтуға болады. Мақаланың шектелуі - оған тек бір алгоритм қолданылған, ал мәліметтер жиынының мөлшері аз.

Түйін сөздер: қазақ тіліндегі көңіл-күйдің талдауы, қазақ мәтіндегі nlp, агглютинативтік тілдер, көңіл-күйдің талдауы, nlp-нің байестік классификаторы.

Introduction

NLP is abbreviated as a natural language processing. It has a lot of applications in the world. In one of the studies, NLP was successfully utilized to detect spam mail in different attachments [1]. It has a vast use in processing of textual information [2]. Generally speaking, whenever we have to work with a text, and we want to somehow apply Artificial Intelligence on it, NLP field has a huge amount of facilities to implement them.

In this work, we decided to make a sentiment analysis due to 2 reasons. The first one is to add contribution to the Kazakh language development in terms of artificial intelligence, and the second one is to make a step for the automatization of evaluation of the university feedback across students in order

to help administrative staff in assessing the current state of quality of education in the university.

Data collection

The data was collected using google forms [3] platform by sharing the link between SDU [4] students. It consists of 2 features and 23 observations (we need an aid of administration to collect more data) :

- opinion about an educational part of SDU;
- whether the opinion is positive or negative.

Out of 23 observations, only 4 of them were negative.

Stopwords are considered as noise in the text of the dataset. In order to remove stopwords, kazakh stopping words have been downloaded from NLTK library and were filtered out from the dataframe. NLTK is a python library, which allows to make lots of NLP operations easier executable [5].

Methods and Results

Explanation of the used algorithm

In AI, gullible Bayes classifiers are a group of basic “probabilistic classifiers” in light of applying Bayes’ hypothesis with solid (innocent) autonomy suppositions between the highlights.

Guileless Bayes has been examined widely since the 1960s. It was presented (however not under that name) into the content recovery network in the mid 1960s, and stays a well known (benchmark) strategy for content order, the issue of making a decision about records as having a place with one classification or the other, (for example, spam or real, sports or legislative issues, and so forth.) with word frequencies as the highlights. With fitting pre-preparing, it is aggressive in this area with further developed techniques including bolster vector machines. It additionally discovers application in programmed restorative conclusion.

Credulous Bayes classifiers are exceptionally adaptable, requiring various parameters straight in the quantity of factors (highlights/indicators) in a learning issue. Greatest probability preparing should be possible by assessing a shut structure articulation, which takes direct time, as opposed to by costly iterative estimation as utilized for some different kinds of classifiers.

In the measurements and software engineering writing, credulous Bayes models are known under an assortment of names, including basic Bayes and autonomy Bayes. Every one of these names reference the utilization of Bayes’ hypothesis in the classifier’s choice standard, however innocent Bayes isn’t (really) a Bayesian technique.

Mathematical formulation of the algorithm

Dynamically, credulous Bayes is a restrictive likelihood model: given an issue occasion to be ordered, spoken to by a vector $\mathbf{x} = (x_1, \ldots, x_n)$ speaking to some n highlights (autonomous factors), it doles out to this case probabilities $p(C_k \mid x_1, \ldots, x_n)$ for each of K possible outcomes or classes C_k.

The issue with the above definition is that if the quantity of highlights n is huge or on the off chance that an element can take on countless qualities, at that point putting together such a model with respect to likelihood tables is infeasible. We in this manner reformulate the model to make it increasingly tractable. Utilizing Bayes' hypothesis, the restrictive likelihood can be disintegrated as

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

In plain English, using Bayesian probability terminology, the above equation can be written as

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

In practice, there is interest only in the numerator of that fraction, because the denominator does not depend on C and the values of the features x_i are given, so that the denominator is effectively constant. The numerator is equivalent to the joint probability model

$$p(C_k, x_1, \ldots, x_n)$$

which can be rewritten as follows, using the chain rule for repeated applications of the definition of conditional probability:

$$\begin{aligned}
p(C_k, x_1, \ldots, x_n) &= p(x_1, \ldots, x_n, C_k) \\
&= p(x_1 \mid x_2, \ldots, x_n, C_k)\, p(x_2, \ldots, x_n, C_k) \\
&= p(x_1 \mid x_2, \ldots, x_n, C_k)\, p(x_2 \mid x_3, \ldots, x_n, C_k)\, p(x_3, \ldots, x_n, C_k) \\
&= \ldots \\
&= p(x_1 \mid x_2, \ldots, x_n, C_k)\, p(x_2 \mid x_3, \ldots, x_n, C_k) \ldots p(x_{n-1} \mid x_n, C_k)\, p(x_n \mid C_k)\, p(C_k)
\end{aligned}$$

Presently the "guileless" restrictive freedom suppositions become an integral factor: expect that all highlights in are commonly autonomous, contingent on the class C_k. Under this assumption,

$$p(x_i \mid x_{i+1}, \ldots, x_n, C_k) = p(x_i \mid C_k)$$

Accordingly, the joint model can be communicated as

$$p(C_k \mid x_1, \dots, x_n) \propto p(C_k, x_1, \dots, x_n)$$
$$= p(C_k)\, p(x_1 \mid C_k)\, p(x_2 \mid C_k)\, p(x_3 \mid C_k) \cdots$$
$$= p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)\,,$$

where $\propto$ signifies proportionality.

This implies under the above autonomy presumptions, the restrictive circulation over the class variable C is:

$$p(C_k \mid x_1, \dots, x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)$$
$$Z = p(\mathbf{x}) = \sum_k p(C_k)\, p(\mathbf{x} \mid C_k)$$

where the evidence is a scaling factor subordinate just $x_1, \dots, x_n$ on that is, a consistent if the estimations of the component factors are known.

Application of the algorithm in research communities

Regardless of the way that the extensive autonomy suspicions are frequently erroneous, the guileless Bayes classifier has a few properties that make it shockingly helpful by and by. Specifically, the decoupling of the class contingent element dispersions implies that every appropriation can be autonomously evaluated as a one-dimensional conveyance. This lightens issues coming from the scourge of dimensionality, for example, the requirement for informational indexes that scale exponentially with the quantity of highlights. While guileless Bayes frequently neglects to create a decent gauge for the right class probabilities, this may not be a necessity for some applications. For instance, the innocent Bayes classifier will settle on the right MAP choice principle arrangement insofar as the right class is more plausible than some other class. This is genuine paying little heed to whether the likelihood gauge is somewhat, or even horribly wrong. As such, the general classifier can be strong enough to disregard genuine inadequacies in its hidden credulous likelihood model. Other purposes behind the watched achievement of the gullible Bayes classifier are talked about in the writing referred to beneath.

Feature extraction

To make the data appropriate to send as an input to the classifier, TF-IDF method has been utilized. It is used for identification of an importance of a word related to a document in a corpus.

Explanation of TF-IDF

In data recovery, tf–idf or TFIDF, short for term recurrence backwards archive recurrence, is a numerical measurement that is proposed to reflect how significant a word is to a record in a gathering or corpus. It is frequently utilized as a weighting element in inquiries of data recovery, content mining, and client displaying. The tf–idf worth expands relatively to the occasions a word shows up in the record and is balanced by the quantity of archives in the corpus that contain the word, which changes for the way that a few words show up more much of the time when all is said in done. tf–idf is one of the most well known term-weighting plans today; 83% of content based recommender frameworks in advanced libraries use tf–idf.

Varieties of the tf–idf weighting plan are regularly utilized via web indexes as a focal device in scoring and positioning a record's significance given a client inquiry. tf–idf can be effectively utilized for stop-words sifting in different subject fields, including content rundown and grouping.

One of the least difficult positioning capacities is processed by adding the tf–idf for each question term; a lot increasingly modern positioning capacities are variations of this straightforward model.

Assume we have a lot of English content archives and wish to rank which report is most pertinent to the inquiry, “the dark colored dairy animals”. A basic method to begin is by wiping out reports that don’t contain every one of the three words “the”, “dark colored”, and “dairy animals”, however this still leaves numerous records. To further recognize them, we may tally the occasions each term happens in each report; the occasions a term happens in a record is called its term recurrence. Be that as it may, for the situation where the length of archives fluctuates enormously, alterations are frequently made (see definition underneath). The principal type of term weighting is because of Hans Peter Luhn (1957) which might be abridged.

The heaviness of a term that happens in an archive is essentially relative to the term recurrence.

Since the expression “the” is so normal, term recurrence will in general inaccurately accentuate archives which happen to utilize “he” more much of the time, without giving enough weight to the more important terms “dark colored” and “dairy animals”. The expression “the” is certifiably not a decent watchword to recognize important and non-pertinent reports and terms, not at all like the less-normal words “dark colored” and “bovine”. Henceforth a reverse record recurrence factor is joined which decreases the heaviness of terms that happen often in the report set and expands the heaviness of terms that happen seldom.

Karen Spärck Jones (1972) considered a factual understanding of term particularity called Inverse Document Frequency (IDF), which turned into a foundation of term weighting.

The particularity of a term can be measured as a reverse capacity of the quantity of records in which it happens.

The tf–idf is the result of two measurements, term recurrence and opposite archive recurrence. There are different ways for deciding the definite estimations of the two insights.

A recipe that expects to characterize the significance of a catchphrase or expression inside a report or a website page.

Various term-weighting plans have gotten from tf–idf. One of them is TF–PDF (Term Frequency * Proportional Document Frequency). TF–PDF was presented in 2001 with regards to recognizing rising subjects in the media. The PDF part estimates the distinction of how frequently a term happens in various spaces. Another derivative is TF–IDuF. In TF–IDuF, idf isn't determined dependent on the archive corpus that will be looked or suggested. Rather, idf is determined on clients' close to home report accumulations. The creators report that TF–IDuF was similarly powerful as tf–idf yet could likewise be applied in circumstances when, e.g., a client displaying framework has no entrance to a worldwide record corpus.

Assume you need to outline a report or a passage utilizing couple of catchphrases.

One strategy is to pick the most much of the time happening terms (words with high term recurrence or tf). Notwithstanding, the most successive word is a less helpful measurement since certain words this way', 'a' happen as often as possible over all records.

Thus, we likewise need a proportion of how interesting a word is for example how rarely the word happens over all records (reverse report recurrence or idf).

Thus, the result of tf x idf (tfidf) of a word gives a result of how continuous this word is in the archive increased by how one of a kind the word is w.r.t. the whole corpus of archives.

Words in the record with a high tfidf score happen much of the time in the report and give the most data about that particular archive.

Basically figuring the recurrence of terms occured in reports experiences a basic issue, all terms are viewed as similarly significant with regards to evaluating importance on an inquiry.

For instance, an accumulation of records on the car business is probably going to have the term auto in pretty much every archive. To end this, an instrument can be presented for lessening the impact of terms that happen time

and again in the accumulation to be important for pertinence assurance. This can be illuminated by downsizing the loads of terms with high accumulation recurrence.

Term Frequency (TF) is the proportion of number of times a word happened in a report to the all out number of words in the record.
For example to represent predisposition against longer records for term recurrence or to represent idf being unclear from a division by zero when the word is absent in the corpus.

Basically figuring the recurrence of terms as in record term lattice experiences a basic issue, all terms are viewed as similarly significant with regards to evaluating pertinence on an inquiry.

For instance, an accumulation of archives on the car business is probably going to have the term auto in pretty much every report. To end this, a system can be presented for lessening the impact of terms that happen time after time in the gathering to be significant for pertinence assurance. This can be comprehended by downsizing the loads of terms with high accumulation recurrence.

Algorithm description

We decided to use MultiBinomial Naive Bayes classifier for the purpose of the paper. The algorithm is suitable in work with discrete values in terms of classification [7].

Results

Given that the class distribution was unbalanced, we decided to show the classification report only of the positive opinion class.

Table 1 : Results of experiment

Precision	Recall	F1 score
0.71	1	0.83

Discussion

Table 1 shows that the model generalizes well and perform in a good manner, currently only more data is needed, and maybe more sophisticated method will be used

Conclusion

As a result, the system of a sentiment analysis identifier on the data about feedbacks across students is working and shows 71% accuracy. It manifests itself as overfitting free and good model given that the data size if very small. But we plan to increase the amount of data with the help of administrative staff of the university and apply more sophisticated model as an

LSTM bidirectional recurrent neural network model to enhance the results and complement them.

References

1 Amol Malge, Dr. S.M. Chaware. An Efficient Framework for Spam Mail Detection in Attachments using NLP. *International Journal of Science and Research (IJSR),* (2016): pp.1121-1125.

2 Arumugam, M. Processing the Textual Information Using Open Natural Language Processing (NLP). *SSRN Electronic Journal,* (2019): pp. 208–217.

3 Google Forms: Free Online Surveys for Personal Use. Google.com. 24 June [Electrnoic resource] URL: https://www.google.com/forms/about/. (Last accessed : 24 June 2019).

4 Suleyman Demirel University - sdu.edu.kz. Sdu.edu.kz [Electronic resrouce]. URL: http://sdu.edu.kz/ (Last accessed : 24 June 2019).

5 Natural Language Toolkit — NLTK 3.4.3 documentation. Nltk.org.. 24 June 2019. [Electrnoic resource] 24 June 2019. URL: https://www.nltk.org/ (Last accessed : 24 June 2019).

6 Rajaraman, A., Ullman, J.D. *Data Mining. Mining of Massive Datasets*, 2011 - pp. 1–17.

7 Scikit-learn.org, 24- Jun- 2019. sklearn.naive_bayes. MultinomialNB — scikit-learn 0.21.2 documentation. 2019.