Big Data Analytics

Lecture 1/B

Supervised classification methods: automatic tagging

UNIVERSITÄT
LUZERN

Reference

- Grimmer, Justin, and Stewart, Brandon M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3): 267-297
- Olivella, Santiago, and Shoub Kelsey (2020). Machine Learning in Political Science: Supervised Learning Models. In Luigi Curini and Robert Franzese (eds.), *SAGE Handbook of Research Methods is Political Science & International Relations*, London, Sage, chapter 56

Classification methods

Supervised Classification Methods

Assigning texts to **some known categories** (rather that to categories *discovered ex-post* the analysis – as it happens with unsupervised classification methods) is the most common use of content analysis methods in social and political science

For example, researchers may ask if local news coverage is positive or negative, if legislation is about the environment or some other issue area, if international statements are belligerent or peaceful, etc.

In each instance, the goal is to infer to which - among a given set of pre-defined categories - each document must be assigned

Classification methods

There are **two broad groups of supervised classification methods** available according to the type of **tagging** (i.e., the assignation of a document to a given pre-defined category) **employed**:

We can have either:

1) human tagging - *supervised learning methods*
2) automatic tagging - *dictionaries*

Human tagging

- ✓ **Supervised learning methods** replicate the familiar manual coding task, *but* with a machine

First, human coders are used to classify a subset of documents into a predetermined categorization scheme (*human tagging*!!!)

Second, this subset is used to train an automated method

Finally, the automated method then classifies the remaining unread documents

- ✓ **Dictionaries** use on the contrary the **relative rate** at which key words appear in a text to **classify documents into categories** (*automatic tagging*! No human intervention in the tagging procedure!)

Let’s first discuss about automatic tagging…

Dictionary methods

Dictionary analysis is very old but still one of the most popular methods in quantitative text analysis for its *technological simplicity*

Suppose the goal is to measure the **tone** (also called the “**sentiment”**) in newspaper articles: whether articles convey information positively or negatively about a given topic or a given politician

What is a **dictionary**?

In this example, a **dictionary to measure sentiment** is simply a list of words that are either dichotomously classified as positive («good», «fantastic», etc.) or negative («bad», «horrible», etc.) or contain more continuous measures (via a set of weights) of their content (to take into account that *fantastic>>good* and *horrible>>bad*)

Dictionary methods

For example, the Lexicoder Sentiment Dictionary (as we will see later in the Lab) aggregates 4,567 sentiment words in the dictionary to two “negative” and “positive” categories

You can then use that dictionary to identify **the tone of a document** (either positive or negative)…according to what?

To the **relative number of words** in that document identified by the dictionary as positive or negative ones!!!

Dictionary methods

Formally, within a given dictionary Z each word *m* (m=1,….M) will have an associated score s_m

For the simplest measures, s_{mn} =1 if the word is associated with a negative sentiment and s_{mp} =1 if associated with a positive sentiment

The analyst then applies some *decision rule*, such as summing over all the weighted feature values, to create a score for the document

Dictionary methods

For example, if $\sum_{m=1}^{M} W_{im}$ are the words included in dictionary Z that are **also** used in document *i*, then you can use a sentiment dictionary to measure the sentiment for any document t_i by developing the following index:

$$t_i = \sum_{m=1}^{M} \frac{s_{mp} W_{im}}{s_{mp} W_{im} + s_{mn} W_{im}}$$

That is, if document *i* presents the words «good», «fantastic» and «bad», then t_i = 2/3 or 0.666

Alternatively, you could also focus on the following sentiment index:

$$t_i = \sum_{m=1}^{M} s_{mp} W_{im} - \sum_{m=1}^{M} s_{mn} W_{im}$$

That is, t_i = 2-1=+1

Dictionary methods

t_i allows therefore to sort documents as to which are more or less positive or negative relative to one other

t_i can also be used to classify documents into **sentiment categories** if a decision rule that identifies a cut point is assumed along with the dictionary method

Perhaps the simplest coding rule would assign all documents with $t_i > 0.5$ (or > 0 according to the index you use) to a positive sentiment category and $t_i < 0.5$ (or < 0) to a negative sentiment

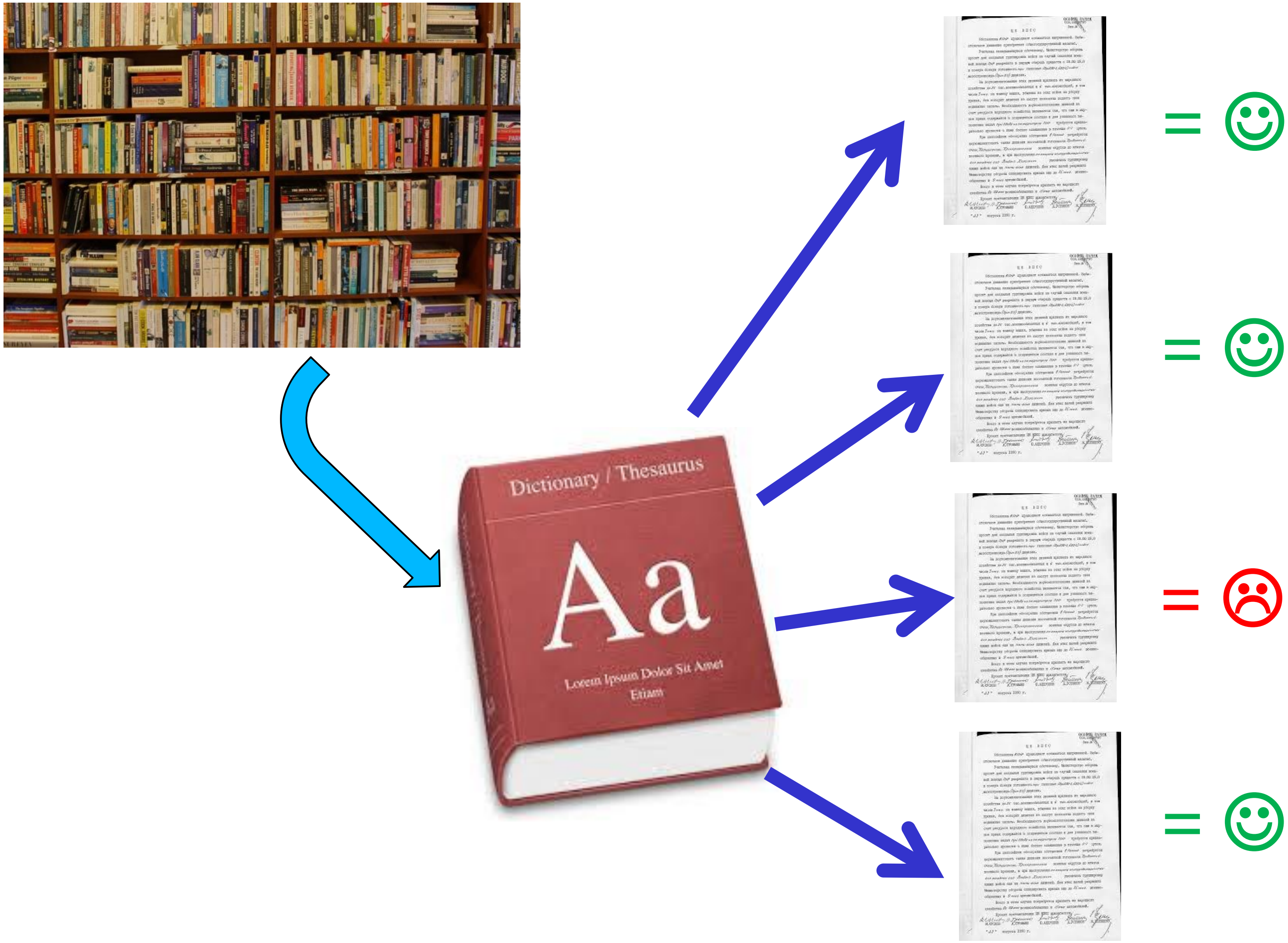
And if $t_i = 0.5$ (or 0)? Either neutral category or NC

Dictionary methods

Of course, the words included in the texts you are analyzing that are **not also included** in the dictionary, will not provide any additional information for your classification aim (we will discuss more about this point later on)

This also explains why the list of words included in a dictionary should be **relatively large**!!!

Dictionary methods



Dictionary methods

Sentiment analysis is just one type of analysis a dictionary method can perform

The general idea of dictionaries is indeed always the same in each given circumstance: identify **words that separate categories** (for example *policy categories*) and measure **how often those words occur in texts**

For example, the Lexicoder Topic Dictionary (Albugh et al., 2013) contains 1,387 keywords under 28 topics (e.g., macroeconomics, civil rights, health care, agriculture) based on the Comparative Agenda Project’s coding scheme. If you are interested about it, just let me know!

Dictionary methods

Using a dictionary not only **minimizes** the amount of labor needed to classify documents (no human involved in the tagging proces after all once a dictionary is already available!)…

…but it also minimizes the computational work of your laptop!

This is very attractive! Once you have for example a sentiment dictionary, you can apply it to any corpus you have **regardless of the size** of the corpus

Dictionary methods

Dictionary methods work pretty well when you study texts that use a **standardized language** (i.e., legal text!).

But it also works good with *noisy corpus* (containing unusual words, tags, etc.), such as a collection of social media posts

This is because the selection of words is based solely on the pre-defined list

Dictionary analysis also works well with *small corpora*, such as responses to open-ended questions, where frequency of words is often too low to perform statistical analysis

Dictionary methods

But…beware of the **challenges** of using a dictionary!

Dictionary methods

First, there is the **problem of availability**!

The very possibility of performing dictionary analysis is dependent on the existence of *suitable dictionaries in the target domain* of your research

Indeed, for dictionary methods to work well, the scores attached to words must closely align with **how** the words are used in a particular context

If a dictionary is developed for a **specific application**, then this assumption should be easy to justify

But when dictionaries are **created in one substantive area and then applied to another**, serious errors can occur

Why that?

Dictionary methods

To build a “good” dictionary you need to be sure that all relevant terms are included in it (**no false negatives**, i.e., terms we should have included in the dictionary cause they are relevant given our research topic, but failed to do so)…

…and no irrelevant or wrong terms are (**no false positives**, i.e., terms we have included in the dictionary but should not have, being them irrelevant given our research topic)

In other words, you want to minimize both false negatives as well as false positives

Dictionary methods

But language do **change across topics**! And when this happens, false negatives and false positives proliferate!

For example, a word like `cancer` may have a positive connotation in a health-care company documents, but negative in many other contexts

Dictionary methods

Note that the possibility of performing dictionary analysis is dependent on the existence of *suitable dictionaries* *also with respect to languages*!

The English language has the largest collection of dictionaries. Several of them also implemented in European languages but not so many in non-European languages

Dictionary methods

A notable effort has been made in recent years to increase the availability of dictionaries in non-English languages using computational tools (such as Google translator: see Proksch et al. 2019)

However, machine translation of a dictionary from English to non-European languages is not always possible because of the absence of one-to-one correspondence of words and the ambiguity of word semantics out of context

Dictionary methods

As a result of this *first problem*, quite often you are called to create your own dictionary from scratch to employ dictionary analysis in new domains or languages

But it is not an easy job to collect usually thousands of words that are related to the target concepts (and that’s why a semi-supervised approach as `newsmap` can be interesting)…

Moreover, you need always to validate your dictionary (more on this later on)

Dictionary methods

Second, there is the problem of the **complexity of language**!

In other contexts, things become more complex…given that **language evolves continuously**: it is a social construction after all!

Dictionary methods

Dictionary methods

That implies, for example, that the list of words included in the dictionary you want to employ (developed at time t-1) can be already (partly) outdated (at time t)…

…**unless you keep updated such list!**

But then several of the advantages of a dictionary begin to fade away!

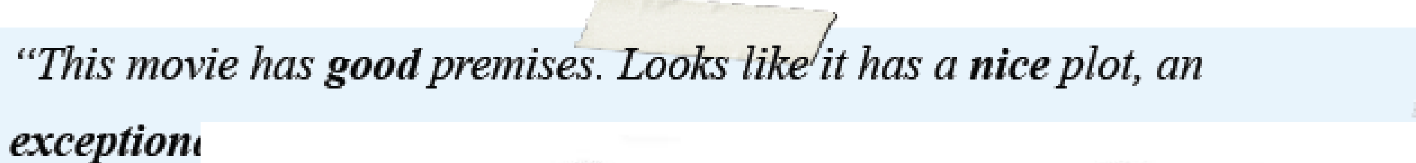
Dictionary methods

Moreover**,** it is almost impossible to code all possible semantic rules in a pre-defined dictionary (double meaning sentences, specific jargons, neologisms, irony)

Dictionary methods

This last point brings us to our third, and final, challenge: the **problem of implementation**!

Counting the number of positive and negative terms in a sentence may lead to **paradoxical effects**



*“This movie has **good** premises. Looks like it has a **nice** plot, an exceptiona*

5 POSITIVE TERMS VS 1 NEGATIVE

AUSTIN COUTURE LUNDGREN LI STALLONE STATHAM CREWS ROURKE WILLIS
THE EXPENDABLES
AUGUST 13

Dictionary methods

Dictionaries, therefore, should be used with **caution**

First: always choose a dictionary **appropriate** to the task at hand

Second: always **validate** the utility of the dictionary, for example by confirming that a sample of dictionary-generated scores of text in the corpus conform to human coding of the text for the measure of interest (i.e., contrast automatic with human tagging)

Dictionary methods

It could be a good idea also to measure the dictionary's reliability to assess the **internal coherence** of the dictionary (i.e., if the included terms are good indicators for the same theoretical concept)

This could be done via a **split-half reliability test**, i.e., 1) we split the dictionary into two parts containing a random half of all dictionary terms; 2) we run the dictionary; 3) we compute the correlation between the results obtained by the two halves of the dictionary and from this the Split-half reliability test

A split-half reliability test helps also to select between competing dictionaries (i.e., you keep the one with the highest outcome for the Split-half reliability test)

Dictionary methods

Summing up: always **avoid** to assume the measures created from a dictionary are correct and then apply them to the problem

The consequence of **domain specificity and lack of validation** can make your analysis built on shaky foundations…