LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

Automatically Selecting Images for News Articles with Keyword Extraction

SYSTEM OVERVIEW

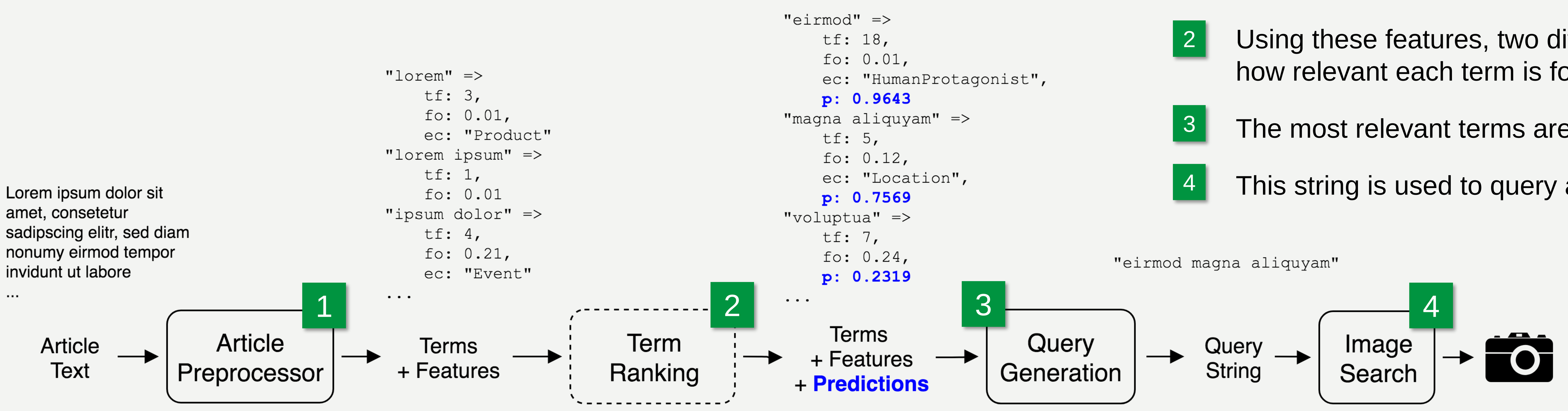


Fig. 1: Overview of the image selection pipeline

1. An **Article Preprocessor** turns the plain text of a news article into a list of terms. Each term is described by certain features.
2. Using these features, two different **ranking mechanisms** predict how relevant each term is for the image search.
3. The most relevant terms are composed into a **query string**.
4. This string is used to query an **image database**.

A. With pre-trained Machine Learning Models

What is a good image search term and what is not? There is no real-world evidence for this, therefore training data had to be assembled from existing information.

When **BBC News** uses photos by **Getty Images**, they sometimes expose the image's ID. I Exploiting this, we downloaded >1500 articles along with image meta data from the Getty database. (see Fig. 2)

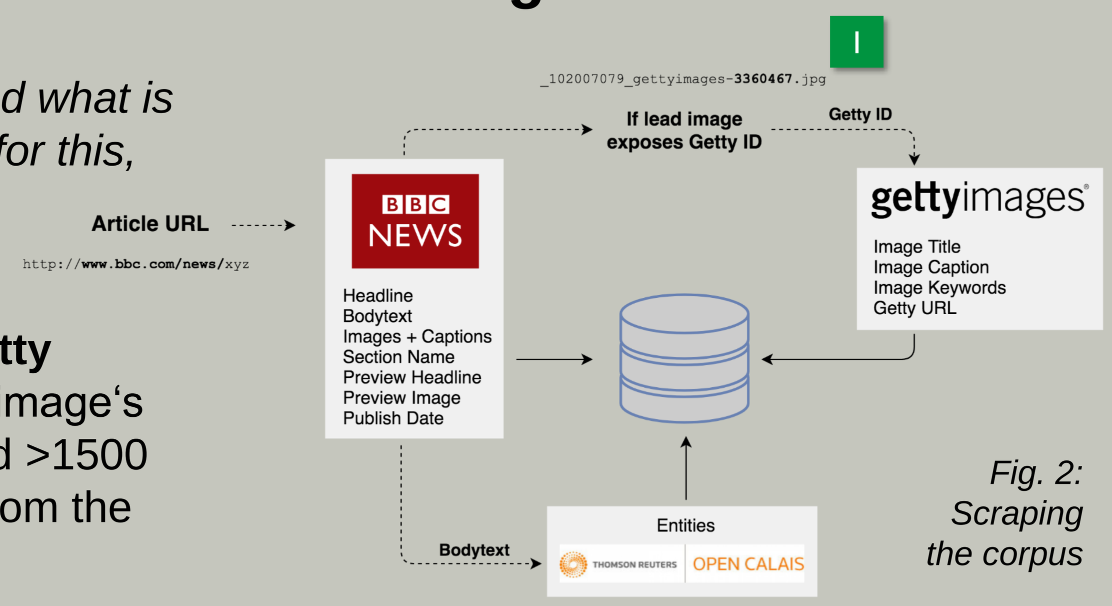


Fig. 2: Scraping the corpus

Assuming that each term in the image meta data suffices as a query to find exactly that image, we generated our training data. Article terms were matched with the image descriptions: Each term that occured both in the article and in the image was labelled as search term. II

From this information, the networks learned which feature values make a good search term.

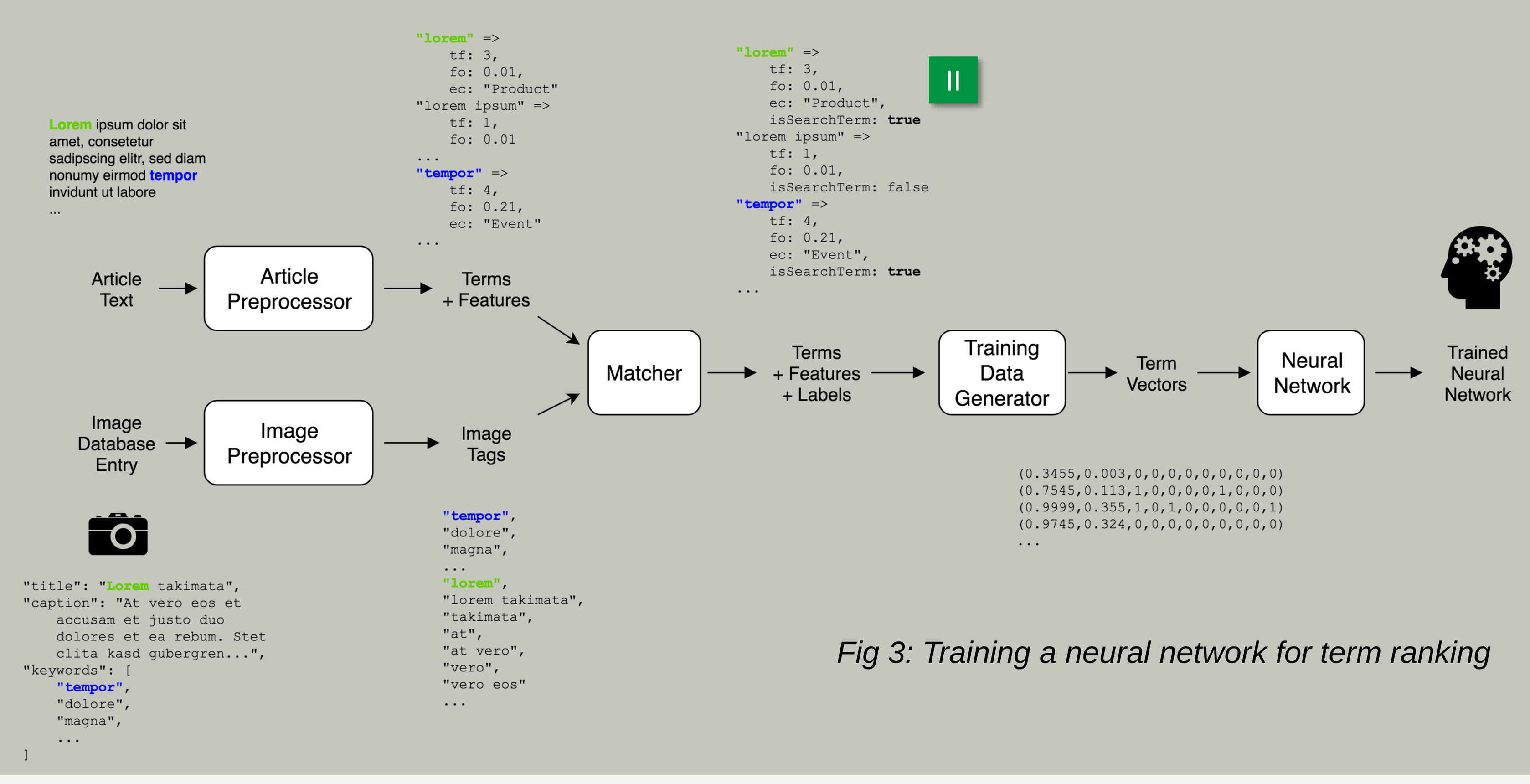


Fig 3: Training a neural network for term ranking

B. With simple Statistics

Assuming that the most relevant terms occur **often and early**, their relevance is calculated as follows:

$$p(tf, fo) = \frac{tf}{\max(TF)} * (1 - fo)$$

where *tf* denotes the frequency of a term, *fo* its first occurrence value and *max(TF)* the highest frequency value of all terms in the article.

Features of a Term

Time waits for no man. Unless that man is Chuck Norris.

Term frequency (tf) is the number of times a term occurs in an article.

tf(„man“) = 2

First occurrence (fo) is the relative position in an article at which a term occurs for the first time – with 0 representing the very first character of the article and 1 the last.

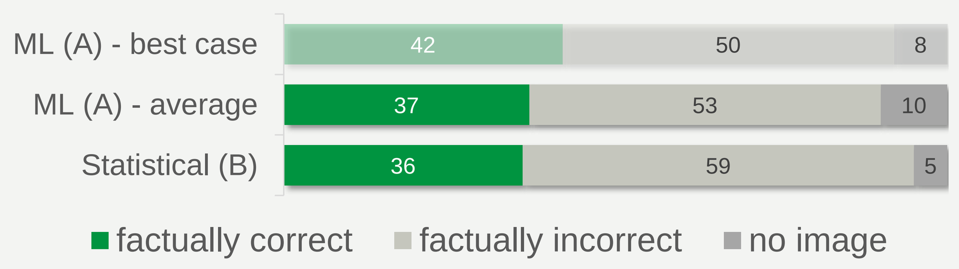
fo(„man“) = 0.3273

Entity category (ec) is a nominal value describing some specific groups of terms, such as „Event“, „HumanProtagonist“ or „Location“.

ec(„Chuck Norris“) = HumanProtagonist

PERFORMANCE

Machine Learning vs. Statistics



a *Fig. 4: Average performance of approaches A and B, including the best performing run for comparison*

Evaluation methodology

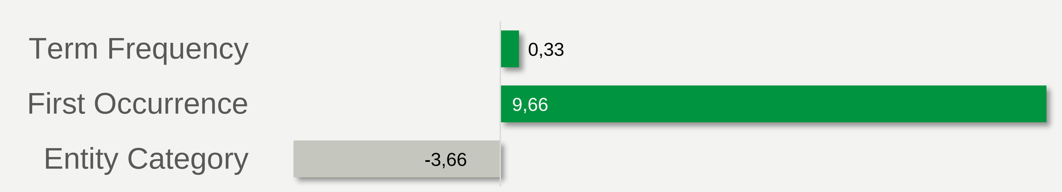
- Sample of 100 BBC articles
- For each article, image selection was run with 4 neural networks + the statistical approach
- Selected images were classified manually as "factually correct" or "factually incorrect", according to our own definition of factual correctness

a The two approaches only differ slightly in their average performance. However, one neural network stood out and selected correct images for 42 percent of all articles.

b Adding the *first occurrence* feature to the neural networks increased the number of correct images by almost 10. In turn, adding *entity category* actually deteriorated the results by almost 4 correct images. *Term frequency* hardly had any impact at all.

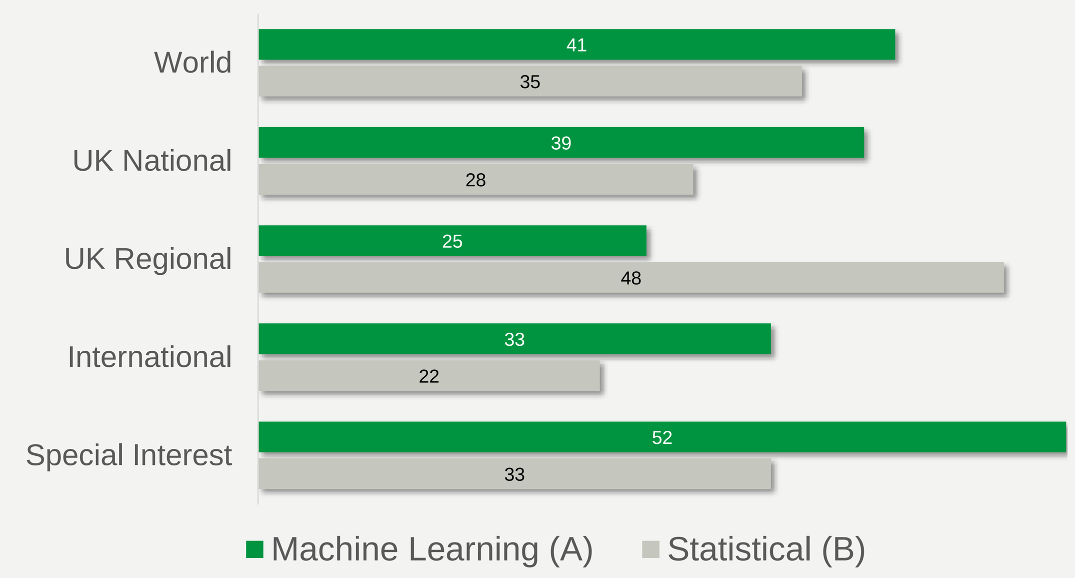
c The Machine Learning approach (A) excelled on articles with special interest topics such as "Entertainment & Arts" or "Business". The more regional the article's scope got, the more did the system's performance decrease – except for the Statistical approach (B) that outperformed the neural networks on regional news.

Impact of the Features



b *Fig. 5: Change in the number of correct images that resulted from adding one specific feature to the neural networks*

Performance per Article Topic



c *Fig. 6: Number of correct images, by article topic and term ranking approach*

Martin Schön
E-Mail: schoen.martin@campus.lmu.de
Phone: +49 172 731 0172

Prof. Dr. Neil Thurman
E-Mail: neil.thurman@ifkw.lmu.de
Phone: +49 89 2180 9449