Olga L.Rink

Moscow State University Institute for Artificial Intelligence

Moscow Institute of Physics and Technology

olga.rink@phystech.edu

Konstantin V.Vorontsov

Moscow State University Institute for Artificial Intelligence

Moscow Institute of Physics and Technology

konstantin.vorontsov@mlsa-iai.ru

Nika K. Shabelskaya

RUDN named after Patrice Lumumba (Moscow)

shabelskaya.nk@mail.ru

**Uncovering positivism, negativism, and conflict in large text collections. Material values and the code of "noble maidens"**

The subject of the sentiments ambiguity expands the scope for inquiring about cultural concepts in large text collections and is another case of applying the “Universal markup” structure and instrument, offered in prior papers.

Keywords: controversy detection, human values, cultural code, content analysis, digital humanity

Among the pitfalls of content analysis within meaning extraction is an observation that opinions communicated in writings are not exclusively positive or negative. A verbal irony when an individual says something but implies its opposite is an example. Labeling a text collection with various tonality classes can help us identify any hidden controversy in social media messages and get more accurate data for sentiment analysis. Thus, introducing the proposed sentiment can make the text markup more efficient.

In prior publications the researchers group offered an approach led into the basis of developing the so called “Universal markup” math and instrument as well as disclosed some pretest statistics for their investigation titled as “Cultural code” [1; 2]. The delivered method of the formalization and the use case of detecting human value concepts in considerable textual collections expand the basic content analysis for a media narrative and is applicable to the ML challenges.

We have uncovered the following human values mostly spread in the data collection under investigation: Material values; Life; Historical memory and continuity of generations; Civic engagement; Political values; Family; Ethnicity; Culture and art; Rights and freedoms; Patriotism; National security; Mutual assistance and mutual respect; Power / Control; Development; Health; Justice; Pacifism; Religiosity; Religious values; Moral values; Dignity; Education; Respect for traditions; Culture (norms) of behavior; Love [1].

The unmarked collection for the trial stage was assembled from the texts of Telegram channels with an audience of more than 5,000 subscribers. 783 documents were published from February 24, 2022 to May, 2022, written in Russian and contained the concept ‘value’ (russ. ‘tsennost’) in any form. Each document was marked up by at least three experts. As a result, 10,426 text elements associated with the value concepts and tonality in 3,496 single annotated documents of the texts were obtained; 1,359 markup elements or 13% of the total amount are annotated as "Controversial Tonality".

TheControversial Tonality label is used in situations where the tonality of a fragment or element linked to a Human value label conflicts with the tonality of the entire text.

1. First we determine the tonality of the entire text relatively to the value in question.
2. Then establish an attitude to the value exclusively at the level of the content of fragments in isolation.
3. Fragments and markup elements whose local values contradict the context of the text in general are marked with the tag "Controversial Tonality".
4. If the text don't represent a clear opinion about the value, but still talks about it, we use a neutral tag.

To conclude, for the subset of annotated documents in the controversial tonality material values turn out to be the most conflicting, while aristocracy (nobility) and some other concepts remain controversy–free cultural codes.

Material values, life, political values, legal awareness (including civic engagement and citizenship), rights and freedoms, and moral values turn out to be the most conflictogenic codes on the corpus of social media documents under investigation. Male orthodoxy, liberalism, political values, and power / control are the foremost disputable in view of conflicting and negative tonality.

On the positive side of the spectrum have been controversy-free aristocracy (nobility), cognitive values, chastity, children, patience, sincerity, aesthetics, modesty, and intelligence.

Every of 105 values in the developed Cultural Code classifier [2] and their groups are represented in the labeled collection and can be utilized for the subsequent ML experimentation.

References

1. Rink O., Lobachev V., Vorontsov K.. Detecting human values and sentiments in large text collections with a context-dependent information markup: a methodology and math. HCII 2024. Lecture Notes in Computer Science series (in print). Cham: Springer.

2. Vorontsov K. V., Gladchenko I. A., Lobachev V. A., Mamontova A. V., Rink O. L., Shabelskaya N. K. Methodology for detecting human values in large text collections // Bulletin of St. Petersburg University. International relations. In Russian (in print).