Context-aware Retrieval and Classification: Design and Benefits

Kurt Englmeier[a]
Faculty of Computer Science, Schmalkalden University of Applied Science, Blechhammer, Schmalkalden, Germany

Keywords: Context Management, Information Extraction, Context-aware Information Retrieval, Named-entity Recognition, Bag of Words, Classification.

Abstract: Context encompasses the classification of a certain environment by its key attributes. It is an abstract representation of a certain data environment. In texts, the context classifies and represents a piece of text in a generalized form. Context can be a recursive construct when summarizing text on a more coarse-grained level. Context-aware information retrieval and classification has many aspects. This paper presents identification and standardization of context on different levels of granularity that supports faster and more precise location of relevant text sections. The prototypical system presented here applies supervised learning for a semi-automatic approach to extract, distil, and standardize data from text. The approach is based on named-entity recognition and simple ontologies for identification and disambiguation of context. Even though the prototype shown here still represents work in progress and demonstrates its potential of information retrieval on different levels of context granularity. The paper presents the application of the prototype in the realm of economic information and hate speech detection.

1 INTRODUCTION

Context-awareness is an important design element of ubiquitous computing (Brown and Jones, 2001). Sensor data, location information, data on user preferences, and the like are gathered, processed, analyzed, and matched in order to compare contexts. The user context, for instance, is compared with the context of her or his surroundings in order to lure her or him into a specific restaurant for lunch. Specific attributes define a context. Attributes such as time of the day, restaurant preferences and actual location of the user may define the context "lunch break opportunities".

In information retrieval, the user query manifests an instance of a user need embedded in its specific context. Because of the representation of the user need being very sparse, systems try to expand the query by suggesting or guessing further query terms. In many cases, query expansion is achieved by observing the behavior of the user community as a whole and gathering common combinations of query terms. Furthermore, search engines often combine query terms with relevant terms from historical data, that is, past queries and selections from retrieval results of the entire user community. The correct interpretation of a user query is pivotal for a successful retrieval of relevant information. However, reasoning the user's information need from a couple of search terms is far from trivial. Producing context information from text is easier. Here, we reflect each statement along the course of a story. Each statement that precedes or succeeds a specific statement contributes valuable information for the correct interpretation of that statement.

Context information can be considered as the product of iterative summarization of statements and standardization of summary terms. This hierarchy of terms constitute semantic anchors of the text on different level of granularity, on phrase or paragraph level or addressing the text in its entirety. The components of the hierarchy, that is, the different semantic anchors, in turn, serve as query expansion.

"Give me all airlines shares that closed yesterday with a loss" replaces cumbersome queries mentioning airlines names and all facets of descriptions of loss.

Separate pieces of text can be linked together to support classification. This can be useful to correctly classify a single piece of text or a statement in a broader context. For example, context information

[a] https://orcid.org/0000-0002-5887-374X

Englmeier, K.
Context-aware Retrieval and Classification: Design and Benefits.
DOI: 10.5220/0009890503020309
In *Proceedings of the 9th International Conference on Data Science, Technology and Applications (DATA 2020)*, pages 302-309
ISBN: 978-989-758-440-4

may classify an apparently innocuous statement as an aggressive and offensive statement when viewed in the broader context of the statements of the same person along a discourse in social media, for example.

Context information helps to focus in and out along generalization and specialization. By generalizing, context information relates, for instance, airline names and their stock market codes to the concept “airline”. In economic analysis, for instance, text analysis must be in the position to recognize all text instances of “output”, “cost”, “lost”, “decrease”, “fraction” and “financing”, just to name a few. This in turn means, that an information retrieval system must resort to concept descriptions that correctly identify all their instances. As a start, we may consider these context descriptions as bag of words containing all terms that specify their respective concept. Furthermore, we combine these terms with named entities in order to address typical expression patterns that stand for a particular concept. This paper presents a system that produces context descriptions from texts in an automatic or semi-automatic way.

In the first phase, we standardize text information as far as possible. We identify different data expressing dates, percentage data, prices, distances and so and annotate them as such resulting in a set of basic named entities. The first phase operates with a number of bag of words (BoW) containing names of locations, countries, etc. It also identifies names from typical patterns, like the key word “Mrs.” or “chancellor” followed by a couple of words starting with a capital letter pointing to a name of a person.

In the next phase, we combine one or two key terms with these basic named entities, looking for 2- or 3-grams containing named entities. We define these patterns of expressions manually. However, the system takes these patterns and tries to find similar patterns, that is, patterns with the same basic named entities but different leading or trailing key words. The identification of a similar pattern in a certain quantity indicates a new instance of the context description.

The patterns identified in this phase are taken as seeds in the next phase of investigating the surroundings of expressions, that is, on a more abstract context level. By repeating this process, we gradually construct a hierarchy of context patterns.

The prototypical system presented here combines named entity recognition (NER) and simple ontologies for the identification of contexts. The paper presents context-aware retrieval and classification in the realm of mining economic texts. The data sources are news articles published by the German Institute of Economic Research (DIW). For this paper, we selected one article from a DIW Weekly Report (Sorge et al., 2020). The economic analysis benefits from context information when a system needs to sift through a large collection of text to find the ones, for instance, that indicate an up- or downswing in a certain industry branch, stock market, or energy consumption.

2 RELATED WORK

Context identification starts with information extraction (Cowie and Lehnert, 1996) and the annotation of the extracted text pieces according to the meaning they express. The annotation is a summary of the extracted text. On a fine-grained level, it is useful to look for patterns that reflect generic information. Such patterns can represent dates, percentages, numerical data, distances, and the like. The combination of factual (numerical) data with text data has its particular appeal. A statistical analysis may come to certain findings. Text mining can help, in parallel, to find statements in articles, news, or Twitter messages that underpin or refute these findings. Numerical analysis, for example, may observe a certain stock by applying time series analysis to measure the probability that its value will rise or drop. Accompanying text analysis sifts through texts and looks for signals that indicate whether this stock is about to take off or drop in value. Identifying these signals and merging them with the numerical analysis rest on quite an array of discovery tasks. Spotting pertinent patterns is quite established in text analysis, in particular in business-related applications, for example in the financial sector (Aydugan and Arslan, 2019).

There are further essential techniques that need to be considered for the design of context-aware retrieval: analysis of word N-grams (Ying et al., 2012), key-phrase identification (Mothe et al., 2018), and linguistic features (Xu et al., 2012; Bollegala et al., 2018; Walkowiak and Malak, 2018). Context-aware retrieval influences also recommender systems and vice versa (Jancsary et al, 2011). The features developed here support the matching of abstract context information and text.

Utterances expressing opinions and, in particular, hate quite often reveal emotions. Hate speech analysis, thus, must consider results and work in textual affect sensing (Liu et al. 2003; Neviarouskaya et al. 2007) alongside discourse analysis. Schneider (2013) developed a framework for narratives of a therapist-patient discourse that is valuable in our context. His work has been summarized and discussed in (Murtagh, 2014).

3 CONTEXT RECOGNITION

3.1 Named Entities at the Basic Level

Information extraction starts with NER of basic and more generic elements referring to time, locations, distances, and the like. This process usually combines key words and patterns of expressions. Finally, it annotates each pattern by an appropriate term that summarizes the meaning of the pattern. The table below indicates a couple of examples of generic patterns.

Table 1: Examples of named entities at the basic level.

Expression	Annotation
between 1979 and 1990	`time span`
by mid-February	`time span`
In the 1950s and ‘60s	`time span`
In July 2019	`date`
123.5	`amount`
25 of the total 30 billion	`fraction`
40 min.	`time span`
850.000	`amount`
six percent	`percentage`
100 kilometers	`distance`

The generic named entities help to standardize factual information and to abstract away the different forms of expressions for essentially the same thing. The examples immediately show (in particular, the second one) that it does not suffice just to annotate the patterns. We save the numerical values in an appropriate way, too. This is the moment when ontologies come into play, because we have to store the numerical information in a suitably standardized way for further interpretation purposes.

3.2 The Role of Bags of Words

NER in the context described here operates with bags of words (BoW) addressing locations, persons, organizations, or institutions (Wall Street, Dow Jones, Casa Blanca, Bangladesh, for instance). Furthermore, we use key words (such as “Mr.” or “Prime Minister”) that hint to names of persons. The system takes these names and feeds them into the respective bag of words. There are further interesting key terms pointing to names. For example, the term “by” following the title of an article leads the list of names authoring this article. The identification of proper names benefits from the analysis of sequential dependencies when bags of words can be produced automatically instead of manually.

There are promising approaches to automatically identify names (and other important key expressions) in texts using conditional random fields (CFR) (Sha and Pereira, 2003) or hidden Markov Models (HMMs) (Freitag and Callum, 2000). Inclined to CFR we integrated a feature that proposes, for example, all names starting with capital letters and followed by an abbreviation as organization names, United Arab Emirates (UAE), or World Nuclear Association (WNA)).

```
<author>
    <person_name>
        Lars Sorge
    </person_name>
    <person_name>
        Claudia Kemfert
    </person_name>
    <person_name>
        Christian von Hirschhausen
    </person_name>
    <person_name>
        Ben Wealer
    </person_name>
</author>

<organization>
    International Atomic Energy Agency (IAEA)
</organization>

<price>
    <money>
        123.5 U.S. dollars
    </money>
    <unit>
        megawatt hour
    </unit>
</price>

<output>
    <power>
        4.8 GW
    </power>
</output>
<distance>
    100 kilometers
</distance>
```

Figure 1: Examples of basic named entities and identified names.

We can easily imagine domain specific BoWs for prices, energy, cooking, travel, and the like. Proper names like the names of persons as shown in figure 1 are fed back to the respective BoWs. Figure 2 shows a couple of examples of named entities extracted from text including names.

3.3 Specific Named Entities

```
<decrease>
    <market>
        Dow Jones
    </market>
    <percentage>
        10,2%
    </percentage>
</decrease>

<decrease>
    <percentage>
        15%
    </percentage>
    <price>
        22.90 dólares
    </price>
    <product>
        barril
    </product>
</decrease>

<investment>
    <production>
        <power>
            5.4 gigawatts (GW)
        </power>
    </production>
    <price>
        28.2 billion U.S. dollars
    </price>
    <price>
        5,300 U.S. dollars
    </price>
    <unit>
        per kilowatt
    </unit>
</investment>
```

Figure 2: Examples of Specific Named Entities.

The next level of abstraction is achieved again by operating on the named entities of the previous phases. Named entities on this level may indicate an increase or decrease in prices, demand, cases, or the like. It may also reflect a current situation on a particular market, country, or industry. Figure 2 shows some examples of specific named entities.

4 CONTEXTUALIZATION ACROSS TEXTS

In social media, we often achieve context-awareness when considering a series of texts in contextual proximity. Statements emerge from events that triggered discourses in diverse social media channels.

4.1 Linking Isolated Statements into Narratives

Hate speech is not isolated or independent from context. It is embedded in the narrative of a person. Her or his narrative joins narratives of further persons constituting a discourse. This discourse, in turn, is rooted specifically in one or more facts emerged or events happened in the past and generally in a socio-cultural context. These sources are in part external to the discourse at hand, but are necessary to correctly interpret meaning and understanding of each utterance in each narrative.

A storyline is a coherent sequence of utterances from mutual narratives that root in things like an event, fact, or statement. It has a timeline that, however, is only of minor importance. Nevertheless, it is time-bound, but only in the sense that its triggering cause happened at a certain point in time. The cause of the discourse (with all its characteristics) and the different persons authoring their respective narratives are the main structural elements of the storyline. The first goal of context-aware classification is to map out the discourse along the storyline. The second goal is to determine heuristics for correctly classifying utterances of hate speech.

The application area presented here is based on a collection of German tweets. It addresses the role and importance of an analysis of statements along the storyline including the anchor texts that triggered the narratives of the storylines.

The sources considered are tweets or comments that, in our example data source, refer to the so-called refugee crisis in Germany, in general, and to specific events with refugees involved. News on such events trail aggressive or offensive comments or posts in newspapers (mostly right-wing ones) and further channels where the news had been re-published. In contrast to traditional media that simply broadcast news, narratives in social media form much more a discourse (or controversy) emerging from the event it is reflecting. News triggering a discourse or controversy has the role of an anchor text.

One of the discourses in our collection, that is used here as an example, rooted in a fatal crime committed by a young refugee that afterwards has been sentenced for murder, and finally committed suicide. The news about this crime is the anchor text, which may be expanded by one or even more news about follow-up events like the conviction and the suicide. The different narratives emerging from that text express the repudiation of the political and justice system in Germany and great parts of the German society. Primarily, they expose a deep and undiscriminating rejection of all refugees, but in particular of these having the same nationality as the young offender. The negative and aggressive narratives also depict a clear picture of the debaters' social anchoring (Meub and Proeger, 2015) that reflects their mental foothold gained from the world view of partisans of right-wing ideology. In that, their anchoring evidences their incapability to make accurate and independent judgements. The following statements are typical for this controversy.

In hate speech detection, it is important to contextualize the discourse over a series of mircoposts. In the end, we want to identify the debater or author of the narrative, target persons or groups, and the debater's leitmotiv (desires, need, and intents) and emotions. To identify the debater's narrative along the storyline is easy. The (real or fake) name of the author is one the few structural elements in tweets and similar messages beside the timestamp. The anchor text can be described using its key terms with or without annotations.

For hate speech detection we apply a particular BoW. containing "toxic" terms (Georgakopoulos et al., 2018) ("fool", "scumbag", "idiot" and the like). Initially, we may consider any occurrence of such a term as toxic, that is potentially discriminating, offensive, or aggressive.

```
<anchor_topic>
    #kandel
</anchor_topic>
<conviction>
    <duration>
        8,5 Jahre
    </duration>
</conviction>
<toxic>
    <duration>
        1,5 Jahre
    </duration>
</toxic>
<affective_state-neg>
    Ich kann gar nicht soviel fresse
</affective_state-neg>
```

Figure 3: Example of a representation of a micropost.

Structural elements like the discourse thread or storyline in which the statement appears and the name of its author are useful for the identification of the statement's context. However, only in rare cases these elements suffice to comprehensively and precisely describe the context. Furthermore, what happens if the statement refers to news or statements outside the storyline? Even if all possible sources of information are within reach, we have to process these sources in order to construct the correct context and to reference to correct things.

The murder of the schoolchild shook the whole of Germany: On December 27, 2017, the Afghan Abdul D. killed his ex-girlfriend Mia (15) with a 20.5 long kitchen knife. On September 3, 2018, he was convicted of insidious murder and assault, was sentenced to eight years and six months.

```
<stabbing>
    <date>
        December 27, 2017
    </date>
    <murder>
        <culprit>
            <person>
                <nationality>
                    Afghan
                </nationality>
                <name>
                    Abdul D.
                </name>
            </person>
        </culprit>
        <victim>
            <person>
                <name>
                    Mia
                </name>
                <age>
                    15
                </age>
            </person>
        </victim>
    </murder>
    <knife>
        20.5 long kitchen knife
    </knife>
</stabbing>
<conviction>
    <date>
        September 3, 2018
    </date>
    <duration>
        eight years and six months
    </duration>
</conviction>
```

Figure 4: Named entity representation of an anchor text in a social media discourse.

By repeatedly applying NER, we standardize and generalize content also across texts. The resulting representations reflect the context of the statements and enables the link to its relevant anchor text.

Let us consider the texts as shown in figure 3 and 4. By the standardized context information in both texts, we are in the position to see that these two texts belong together. Furthermore, with the overall context information we can classify the text of figure 4 as hate speech.

4.2 Phases of Contextualization in Hate Speech Detection

Hate speech-related text features are probably best detected along a supervised learning process (Chatzakou et al., 2019). Our system supports hate speech detection over a series of phases. In each phase it applies NER as outlined above together with bag of words.

1. Identifying structural elements of the discourse, its time frame, anchor text, and the different narratives of the debaters.
2. Cleansing obfuscated expressions, misspellings, typos and abbreviations by applying character patterns and distance metrics.
3. Application of different bags of words to locate mentions of persons, groups, locations etc.
4. Identifying outright discriminating, offensive, and aggressive terms.
5. Identifying emotions and measuring the affective state.
6. Measuring the toxicity of individual statements and narratives.

The process of phase 1 yields a linked list containing the individual statements with their time stamps and pointer to its author and anchor text.

Phase 2: The next step, the cleansing process, addresses terms that are intentionally or unintentionally misspelled or strangely abbreviated:

- “@ss”, “sh1t”, “glch 1ns feu er d@mit”, correct spelling: “gleich ins Feuer damit”: “[throw him/her/them] immediately into the fire”.
- “Wie lange darf der Dr*** hier noch morden?”: “How long may this sc*** still murder? “Dr***” stands for “Drecksack (scumbag)”.

Phase 3: Contexter uses here bags of words containing names of persons, locations, prominent groups, parties, and the like (including synonyms), even though there exist promising approaches for automatically identifying names of in texts based on conditional random fields, for instance (Sutton and McCallum, 2012).

Phase 4: Further bags of words contain toxic terms. The toxicity is approved if no immediate negation reverses the polarity of the expression.

The example of figure 5 shows how two potentially toxic expressions turn the statement into an aggressive one. The close proximity of the toxic expression to the threat, that is, with only (presumably) profane expressions in between, clearly indicates the author’s wish to do severe harm to politicians. This conclusion can be achieved by the system in an automatic way. The schema works also for similar mentions when different targets addressed like a religious group, a minority, or a prominent person in conjunction with a threat. The example also shows some typical misspellings or intentional typing errors.

```
Das tolle land haben nur die
<aggression><toxic><korrupt>kuropten</korrupt>
<target>Politiker</target></toxic> die ieben in saus und braus.
Nur die kleinen müssen zusehen wo sie bleiben. glch
<toxic><ins Feuer>1ns feu er</ins Feuer>
d@mit</toxic></aggression>
```

Figure 5: The potentially toxic expression (“corrupt politicians”) turns the initially profane expression (“into the fire”) into an aggressive statement.

The tweet of figure 5 can be classified as hate speech even without consideration of the preceding storyline the tweet is part of. However, there are cases when we need background information. Imagine the statement “send them by freight train to …” instead of “into the fire”. “Freight train” in the context of hate speech has always a connotation with the holocaust. The cruelties of the Nazi regime provide important background information, we have to take into account in hate speech analysis. This background is just as important as the anchor text.

```
<anchor topic>#kandel</anchor topic>
<aggression><conviction>8,5 Jahre
Jugendstrafe</conviction> für einen MORD!
<toxic>Wofür hab es die 1,5 Jahre Rabatt???</toxic>
<affective state-neg>Ich kann gar nicht soviel fressen,
wie ich kotzen möchte</affective
state-neg></aggression>
```

Figure 6: Example of an expression of a negative affective state expressed in “statement 1”.

Phase 5: In hate speech, we encounter many expressions of positive or negative emotions. These expressions are an important indicator of the overall affective state of the author in relationship to the discourse or the facts as described in the anchor text. The last phrase in figure 6 (“I can't eat as much as I want to puke.”) insinuates a negative affective state of the author. The reference to the anchor text addressing the details of this event is important for the correct classification of this tweet. The anchor text ("Kandel") provides information on the crime of the young offender and his conviction. The close proximity of the fact to the author's negative affective state reveals her or his repudiation of the conviction. We may take this affective state as a special indicator that has a negative impact on its surrounding, which can be toxic statements or facts from the anchor text or the immediate statements from the other debaters.

Phase 6: The final measurement of the toxicity combines the evaluations obtained from individual statements with related affective states.

The measurement of the toxicity depends on the quantity and quality of aggressive terms in the statement. Here, our System differentiates between oppositional opinion, offensive statement, threat against something or somebody, or inciting statement. In some cases, qualification is straightforward. For example, if the author of the statement uses outright aggressive terms like in "Ich bin dafür, dass wir die Gaskammern wieder öffnen und die ganz Brut da reinstecken.- I'm in favor of opening the gas chambers again and put in the whole offspring.", we can immediately classify this statement as hate speech. In all other cases, we combine the levels of toxicity assigned to that statement. The overall scenario, for instance, may simply be an oppositional opinion. However, combined with a strong negative affective state (similar to one of Statement 1) the statement as whole qualifies as offensive statement. For the time being, our system evaluates each statement independently. However, in the near future it will try to capture the latent prevailing mood or opinion of the author along her or his narratives.

5 CONCLUSIONS

This paper presented the state of work of a prototypical system to produce and apply context-aware information retrieval and classification on different levels on granularity. Named entity recognition (accompanied by analysis of N-grams) helps to identify context information.

The paper presents application of recursive NER in the area of economic analysis and hate speech detection. Once the context descriptions are created, retrieval and classification processes operate on these data. It enables a smoother navigation over texts and zooming in to text passages that hit the interest of the users. It supports also the contextualization across a series of statements along their discourse storyline in social media. Text analysis along the storyline of discourses supports hate speech detection.

The long-term objective of the system design as discussed here is a stronger involvement of humans in the development of context information and on the behavior of the system concerning context inference. This involvement results in a more active role of the users in designing, controlling, and adapting of the learning process that feeds the automatic detection of context information.

REFERENCES

Aydugan Baydar G. and Arslan S. (2019). FOCA: A System for Classification, Digitalization and Information Retrieval of Trial Balance Documents.In *Proceedings of the 8th International Conference on Data Science, Technology and Applications – DATA*, pp. 174-181.

Bedathur, S., Berberich, K., Dittrich, J., Mamoulis, N., Weikum, G., 2010. Interesting-phrase mining for ad-hoc text analytics. In: *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, 1348-1357.

Bollegala, D., Atanasov, V., Maehara, T., Kawarabayashi, K.-I., 2018. ClassiNet—Predicting Missing Features for Short-Text Classification. *ACM Transactions on Knowledge Discovery from Data (TKDD*) 12(5): 1–29.

Brown, P.J., Jones, G.J.F., 2001. Context-aware Retrieval: Exploring a New Environment for Information Retrieval and Information Filtering. In: *Personal and Ubiquitous Computing* 5(4): 253–263

Chatzakou, D., Leontiadis, I., Blackburn, J., Cristofaro, E. de, Stringhini, G., Vakali, A., Kourtellis, N. (2019). Detecting Cyberbulling and Cyberaggression in Social Media. *ACM Transactions on the Web (TWEB)* 13(3), pp. 1–51.

Cowie, J., Lehnert, W., 1996. Information Extraction. *Communications of the ACM*, vol. 39, no. 1, 80-91.

Fan, W., Wallace, L., Rich, S., Zhang, Z., 2006. Tapping the power of text mining. *Communications of the ACM*, vol. 49, no. 9, 76-82.

Freitag, D., McCallum, A., 2000. Information Extraction with HMM Structures Learned by Stochastic Optimization. *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence,* pp. 584-589.

Georgakopoulos, S.V., Tasoulis, S.K., Vrahatis, A.G., Plagianakos, V.P., 2018. Convolutional Neural Networks for Toxic Comment Classification. In: *Proceedings of the 10th Hellenic Conference on Artificial Intelligence,* pp. 1–6.

Jancsary, J., Neubarth, F., Schreitter, S., Trost, H., 2011. Towards a context-sensitive online newspaper. In: *Proceedings of the 2011 Workshop on Context-awareness in Retrieval and Recommendation*, pp 2-9.

Laborde, A., 2020. Wall Street cae un 6,3% en una sesión

en que llegó a retroceder el 10%. *El Pais*, March 18, 2020.
Liu, H., Lieberman, H., Selker, T. (2003). A model of textual affect sensing using real-world knowledge. In: *Proceedings of the 8th international conference on Intelligent user interfaces,* pp 125-132.
Meub, L., Proeger, T.E. (2015). Anchoring in Social Context, *Journal of Behavioral and Experimental Economics* (55), pp. 29-39.
Mothe, J., Ramiandrisoa, F., and Rasolomanana, M. (2018). Automatic Keyphrase Extraction Using Graph-based Methods. In: *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, pp. 728–730.
Murtagh, F., 2014. Mathematical Representations of Matte Blanco's Bi-Logic, based on Metric Space and Ultrametric or Hierarchical Topology: Towards Practical Application. *Language and Psychanalysis*, 3 (2), pp. 40-63.
Neviarouskaya A., Prendinger H., Ishizuka M., 2007. Textual Affect Sensing for Sociable and Expressive Online Communication. In: Paiva A.C.R., Prada R., Picard R.W. (eds) *Affective Computing and Intelligent Interaction. ACII 2007. Lecture Notes in Computer Science*, vol. 4738.
Sha, F., Pereira, F., 2003. Shallow Parsing with Conditional Random Fields. *Proceedings of the HLT-NAACL conference*, pp. 134-141.
Schneider, P., 2013. Language usage and social action in the psychoanalytic encounter: discourse analysis of a therapy session fragment. *Language and Psychoanalysis*, 2 (1), pp. 4-19.
Sorge, L., Kemfert, C., von Hirschhausen, C., Wealer, B., 2020. Nuclear Power Worldwide: Development Plans in Newcomer Countries Negligible. *DIW Weekly Report* 10 (11), pp. 164-172.
Sutton, C., McCallum, A., 2012. An Introduction to Conditional Random Fields, Foundations and Trends in *Machine Learning* 4(4), pp. 267–373.
Walkowiak T. and Malak P. (2018). Polish Texts Topic Classification Evaluation.In Proceedings of the 10th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART, pp. 515-522.
Xu, J.-M., Jun, K.-S., Zhu, X., Bellmore, A., 2012. Learning from bullying traces in social media. In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 656–666.
Ying, Y., Zhou, Y., Zhu, S., Xu, H., 2012. Detecting offensive language in social media to protect adolescent online safety. In: *Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust, PASSAT 2012, and the 2012 International Conference on Social Computing, SocialCom 2012*, pp. 71-80.