Who likes me more? Analysing entity-centric language-specific bias in multilingual Wikipedia

Yiwei Zhou
Department of Computer Science
University of Warwick
Coventry, UK
Yiwei.Zhou@warwick.ac.uk

Elena Demidova
L3S Research Center
Hannover, Germany
demidova@L3S.de

Alexandra I. Cristea
Department of Computer Science
University of Warwick
Coventry, UK
A.I.Cristea@warwick.ac.uk

ABSTRACT
In this paper we take an important step towards better understanding the existence and extent of entity-centric language-specific bias in multilingual Wikipedia, and any deviation from its targeted neutral point of view. We propose a methodology using sentiment analysis techniques to systematically extract the variations in sentiments associated with real-world entities in different language editions of Wikipedia, illustrated with a case study of five Wikipedia language editions and a set of target entities from four categories.

CCS Concepts
●Information systems → Web mining; ●Computing methodologies → Natural language processing;

Keywords
Multilingual Wikipedia; Sentiment Analysis; Linguistic Point of View

1. INTRODUCTION
Over the recent years Wikipedia has expanded into a large and much used source of information on the Web (with almost 24 million users, and growing at a rate of 10 edits/sec by editors from all around the world [2]). Thus, Wikipedia is currently available in more than 280 different language editions [1] that are being increasingly interlinked. As such, Wikipedia has become a valuable cross-lingual information source. However, as different language editions of Wikipedia evolve independently, semantic differences between the language-specific editions of the articles may occur.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SAC 2016, April 04 - 08, 2016, Pisa, Italy
Copyright 2016 ACM 978-1-4503-3739-7/16/04...$15.00
http://dx.doi.org/10.1145/2851613.2851858

For example, in [17], the author illustrated on one example that, although Wikipedia aimed at the Neutral Point of View (NPOV), such NPOV can vary across its language editions, building linguistic points of view (LPOV).

For a more systematic approach, sentiment analysis is an important technique that is able to automatically quantify and thus better understand bias in multilingual Wikipedia and differences in the representations of specific entities across different language editions. However, very few studies considered sentiment analysis on multilingual text collections [8]. Moreover, existing sentiment analysis techniques mostly focus on document collections from domains with explicit sentiment expressing purpose, and often having a clear structure (e.g., product reviews). Given its NPOV aim, such existing tools are not directly applicable to determine language-specific bias in an encyclopaedia like Wikipedia, where we expect much more moderate differences, which we aim at capturing.

An important limitation of the existing studies on multilingual Wikipedia is their focus on the comparative analysis of one entity-related article at a time (e.g. [17] and [11]). However, even a dedicated Wikipedia article can typically cover only a part of the facts associated with an entity in Wikipedia and thus cannot fully reflect the language-specific bias associated with this entity. Although an exhaustive collection of all mentions of every entity in Wikipedia does not appear feasible, due to the size and the constant growth of the dataset, entity occurrences across articles are often reflected in the Wikipedia link structure that can be effectively analysed using the methods proposed in this paper.

Therefore, here we take a first and important step towards better understanding language-specific sentiment bias in entity representation in multilingual Wikipedia. First, we propose a methodology to systematically generate entity-centric graphs that cover multiple occurrences of the entity across the articles in a Wikipedia edition. Second, we analyse the differences in the sentiments associated with real-world entities in different language editions of Wikipedia. Third, we apply our techniques in a case study encompassing five Wikipedia language editions, analyse and discuss the differences in sentiments with respect to the entities under consideration.

The results of the case study applied to more than 1,000,000 sentences containing the entities under consideration illus-
trate that, although the majority of content in Wikipedia is
obeying the NPOV principle, a moderate but stable amount
of sentiment-expressing information (around 8% in average,
but differs from entity to entity) is to be found in every
language edition, representing positive as well as negative
sentiments; importantly, these sentiments, and the entities
they refer to, are often language-specific.

The rest of the paper is organised as follows: Section 2
presents the methodology of our study including a descrip-
tion of the processing pipeline to create entity-centric graphs
from multilingual Wikipedia using link structure and annotation
tools. Then, in Section 3 we present the
results of a case study applying the proposed methodology
to five Wikipedia language editions and 219 target entities
from four different categories. Following that, in Section 4
we discuss related studies on multilingual Wikipedia, senti-
ment analysis and entity extraction. Finally, we discuss our
results and provide a conclusion in Section 5.

2. ENTITY-CENTRIC SENTIMENT ANAL-
YSIS OF MULTILINGUAL WIKIPEDIA

The proposed easily-reproducible processing pipeline for
analysing multilingual Wikipedia articles (presented in Fig-
ure 1) contains the following steps:

1. Article Extraction: In this step we use the Wikipedia
   link graph including in-links and language links to se-
   lect the articles that are likely to contain references to
   the target entity.

2. Sentence Extraction: To reduce the amount of
   the plain text to be processed further, we preselect the
   sentences that are likely to mention the target entity
   using a dictionary-based method.

3. Sentence Translation: To enable homogeneous pro-
   cessing of multilingual Wikipedia sentences, we trans-
   late the extracted sentences to English using machine
   translation.

4. Entity Annotation: To disambiguate entity surface
   forms in the translated sentences, we use the DBpedia
   Spotlight service. This step helps to filter out the
   sentences that contain ambiguous entity surface forms
   identified in Step 2 that are not relevant to the target
   entity.

5. Sentiment Analysis: We apply sentiment analysis tech-
   niques and analyse the aggregated sentiment scores of
   the sentences remaining after the filtering in Step 4.

6. Result Analysis: At the end of the process, the entity-
   centric data from the multilingual Wikipedia articles
   is annotated and available for further analytics.

In the following we describe these steps in more detail.

2.1 Article Extraction

The first step in the pipeline is to collect the articles that
are likely to mention the target entity. As currently there
are more than four million articles alone in the English
Wikipedia, and only a few of them are relevant to any
specific entity, it is not efficient to analyse all the articles
for each target entity. Therefore, we automatically analyse
the link structure of Wikipedia, to narrow down the search
space. In particular, we use in-links of the main Wikipedia
article representing the target entity (i.e., other articles that
have a link to this page), as well as the language links of these
articles (i.e., Wikipedia articles on the same entity in other
Wikipedia language editions). To access the link structure
of the Wikipedia articles, including the in-links and the lan-
guage links, we use the MediaWiki API.

Figure 2 illustrates our Article Extraction method, by show-
ing an example of how we extract the articles that are
likely to contain information on GlaxoSmithKline (a British
healthcare company) — one of our target entities from the
German and English Wikipedia. In Figure 2, we use small
rectangles to represent different articles, and different fill
patterns of the rectangles to represent the languages of
these articles. For the German Wikipedia, we build a set
of articles in four steps: First, the article in the German
Wikipedia describing GlaxoSmithKline. Second, all other
articles from the German Wikipedia, which are linked to
this Wikipedia article. Most of these articles are in Ger-
man, e.g. DTP-Impfstoff (DTP-vaccine in German), AT&T,
Chlorambucil, Sage Group, etc. Third, for all the articles
from other Wikipedia language editions describing Glaxo-
SmithKline (we use English in Figure 2 as an example), we
extract all the articles linked to them. Most of these arti-
cles are in English, e.g. DTP-vaccine, Chlorambucil, Sage
Group, Beckman Coulter, etc. Fourth, if one of these arti-
cles, such as the article describing Beckman Coulter in Fig-
ure 2, also has a German edition that has not yet been added
to the German Wikipedia article set on GlaxoSmithKline,
our algorithm adds that article to the article set. After nar-
rowing down the search space using this procedure, we re-
trieve the content of these articles (that are most likely
to mention the target entity) through the MediaWiki API.

2.2 Sentence Extraction

Wikipedia articles are of different lengths, often containing
dozens of sentences. However, only few of these sentences
are relevant for any target entity. Moreover, machine translation
(applied in the following steps) is an expensive procedure,
which requires external services that normally restrict the
number of requests or the amount of translated text. Also
entity disambiguation services, such as DBpedia Spotlight,
require API requests and bring in latency issues. There-
fore, it is desirable to pre-select relevant sentences that are
likely to mention the target entity before further processing
steps, and especially machine translation and entity disam-
biguation, are performed. Although entity annotation tools
for English are well-developed, that does not apply to many
other languages. Therefore, to facilitate pre-selection of rele-
vant sentences, in this step we perform rough entity annota-
tion in multilingual Wikipedia articles (without translating
them) using a dictionary-based method.

The articles obtained directly from the MediaWiki API con-
tain metadata carrying no sentiment information, such as
HTML tags, references and sub-titles. Therefore, first of all, we eliminate these metadata to obtain plain text for

http://www.mediawiki.org/wiki/API:Main_page
each selected Wikipedia article. Then, we retrieve possible surface forms of the entity in the target languages using DBpedia\(^2\). DBpedia contains structured information extracted from many Wikipedia language editions. The target entity can have different surface forms in the Wikipedia articles. For example, for the entity “Angela Merkel”, the Chancellor of Germany, her corresponding English DBpedia entry is [http://dbpedia.org/page/Angela_Merkel](http://dbpedia.org/page/Angela_Merkel). From this entry, we can find her names, alias and titles in English referring to different time periods, such as “Angela Merkel”, “Angela Dorothea Kasner”, “Chancellor of Germany” and “Angela Dorothea Merkel”. Besides that, she might also be referred to as “Angela”, “Merkel”, “Chancellor Angela Merkel” or “Ms. Merkel” in different Wikipedia articles.

After the surface forms of an entity for a language are obtained from DBpedia, we search for these surface forms in the articles extracted from the corresponding Wikipedia language edition and extract the sentences containing these forms. To extract the relevant sentences surrounding the identified entity surface forms we employ the NLTK\(^3\) sentence tokeniser. This tokeniser was trained on and worked well for many European languages \([12]\), being able to segment articles into sentences correctly when facing sentence separators from different European languages. Other sentences, which do not mention any relevant surface form, are discarded.

This procedure significantly reduces the volume of text to be analysed further, resulting in a lower number of requests to machine translation and entity disambiguation services in the following steps and overall efficiency of the proposed method. The result of the Sentence Extraction step is a set of sentences in different languages that contain the predefined surface forms of the target entity.

It should be noted that even if a sentence contains such predefined entity surface forms, it does not always mean that this sentence is relevant to the target entity due to the ambiguity of surface forms. Therefore, we apply entity disambiguation to the pre-selected sentences later in the pipeline, as described in Section 2.4.

Even though this step inevitably discards some sentences using anaphoras referring to the target entity, it can achieve really high precision and efficiency combined with the Entity Annotation step.

### 2.3 Sentence Translation

For a fair comparison, we would need sentiment analysers that have the same performance on different languages, which is not feasible to achieve. Thus, instead, in order to bring multilingual text to a common denominator, we translate all the non-English sentences remaining from the last step to English, using automatic translation methods, which enables us to use the same English sentiment analysis resources to measure the sentiment strength of the multi-
lingual text. Nowadays, machine translation has become a mature technique, which is widely used in research and business, and multiple translation tools have been released by different organisations. Among them, we selected Google Translate\footnote{https://translate.google.co.uk/} for its good accuracy and usage history in the multilingual sentiment analysis area. For example, Wan \cite{wan2011labeled} used Google Translate to close the gap between an English training data set and Chinese test data set; Banea et al. \cite{banea2010machine} employed Google Translate on Romanian and Spanish text to use English subjectivity analysis resources on them. Their research results have shown the effectiveness of this machine translation service. Besides, in the later steps, we apply lexicon-based sentiment analysis techniques, so that grammatical errors that could have been introduced during the translation step will not affect the sentiment analysis step. This is an additional precautionary measure, as the lexicon-based method that we use is expected to introduce very few — if any — errors with the translation, unlike the context-dependent sentiment analysing techniques.

### 2.4 Entity Annotation

Even through now all the translated sentences contain at least one of the predefined entity surface forms, not all of them are relevant to the target entity. For example, an occurrence of the surface form “Merkel” can be ambiguous. There exist a town in the United States and a motorcycle company that are also named “Merkel”. As these examples illustrate, precise entity extraction and disambiguation tools play an important role in creation of the entity-centric graph over the multilingual Wikipedia. Therefore, we take a further step to eliminate potentially irrelevant sentences. The aim of the Entity Annotation step is to disambiguate the surface forms of the target entity in the translated sentences. DBpedia Spotlight \cite{flesca2013spotlight} is a system that can automatically annotate text documents with DBpedia URIs. It uses the DBpedia Lexicalisation dataset to provide candidate disambiguations for each surface form in the text, and a vector space model to find the most likely disambiguation \cite{flesca2013spotlight}. It achieves a competitive performance with other commercial tools, is configurable and free for the public to use. Due to the above reasons, we select DBpedia Spotlight as the annotation tool. To balance precision and recall in the annotation process, we experimentally set the confidence threshold of DBpedia Spotlight to 0.6 (we also experimented with the thresholds of 0.2, 0.5, and 0.8, whereas 0.6 achieved the best performance with respect to the $F_1$ measure). After the annotation with DBpedia Spotlight, all the mentions of the target entity in each sentence are annotated. We discard the sentences without any annotations of the target entity. After this step, we obtain the set of sentences that are relevant to the target entity.

### 2.5 Sentiment Analysis

The aim of the Sentiment Analysis step is to annotate each sentence with the sentiment strength towards the target entity. We are interested in the aggregated sentiment of Wikipedia rather than the sentiments of separate sentences. As illustrated in \cite{truyen2012multilingual}, for the lexicon-based method, when facing a fairly large number of sentences, the errors in polarity detection will cancel out relative to the quality we are interested in analysing. Besides that, in \cite{truyen2012multilingual}, researchers pointed out for methods with high performance on individual sentences, when applied to analyse a large number of sentences, the results could be largely biased.

In order to enable homogeneous processing of entities from different domains and languages, obtain aggregated and graded sentiment strength scores, we apply a lexicon-based method. To this extent, we employ SentiWordNet \cite{esnas2014multi}: a state-of-the-art lexicon containing more than 20,000 sentiment bearing words, being the sentiment lexicon with the widest coverage to date. SentiWordNet annotated all the words in WordNet\footnote{http://wordnet.princeton.edu} with three numerical scores (adding to 1): positive, negative and objective, each in the range of 0 to 1. Many researches, for example, \cite{truyen2012multilingual, ma2013 Sentiment, liu2013towards}, utilised SentiWordNet to analyse sentiment polarity and strength.

To enable sentiment analysis, we tokenise each sentence in our final sentence set with the Natural Language Toolkit (NLTK\footnote{http://www.nltk.org}), then parse it with the Stanford Part-Of-Speech Tagger\footnote{http://nlp.stanford.edu/software/tagger.shtml}. The Stanford POS tagger achieved a 97.24% accuracy on WSJ \cite{manning2014stanford}, a dataset with formal sentences, similar to sentences in Wikipedia articles. After getting the POS tag, we lemmatise each word with NLTK and use the lemmatised words and POS tags to obtain the positive, negative and objective sentiment scores from SentiWordNet.

At the sentence level, we aggregate the sentiment scores of the sentiment-bearing words in the sentence. To eliminate the influence of the length differences among sentences, we normalise the resulting sentence sentiment scores by the number of sentiment bearing words in this sentence (i.e., the words matched with SentiWordNet). In summary, the positive, negative and objective sentence scores $S(s_j, sentiment)$ of a sentence $s_j$ towards the target entity contained in this sentence are calculated as follows:

$$S(s_j, sentiment) = \frac{\sum_{i=1}^{m} S(w_i, sentiment)}{m}.$$  

where: $w_i$ is the $i^{th}$ sentiment bearing word that occurs in the sentence $s_j$ and in SentiWordNet; $S(w_i, sentiment)$ represents the sentiment score (positive, negative or objective) of the word $w_i$; $m$ is the total number of sentiment bearing words in $s_j$.

The numbers of sentences extracted for a given entity from different language editions vary. Therefore, to make the sentiment scores comparable across different language editions, we need to further normalise the sentiment scores, by taking into account the number of sentences extracted from the language edition. To this extent, we build average positive, negative and objective scores per sentence in a language, for each target entity.

The positive, negative and objective sentiment scores $S(l, sentiment)$ for a language $l$ towards the target entity

\begin{align}
\text{positive score:} & \quad \frac{1}{n_l} \sum_{s_j \in \text{sentences}_{l}} S(s_j, sentiment) \\
\text{negative score:} & \quad \frac{1}{n_l} \sum_{s_j \in \text{sentences}_{l}} -S(s_j, sentiment) \\
\text{objective score:} & \quad \frac{1}{n_l} \sum_{s_j \in \text{sentences}_{l}} |S(s_j, sentiment)|
\end{align}
are calculated, respectively, as follows:

\[ S(l, \text{sentiment}) = \frac{\sum_{j=1}^{n} S(s_j, \text{sentiment})}{n} \tag{2} \]

where: \( s_j \) is the \( j^{th} \) sentence in language \( l \) that mentions the target entity; \( S(s_j, \text{sentiment}) \) represents the sentiment (positive, negative or objective) score of the sentence \( s_j \); \( n \) is the total number of sentences that mention the target entity in \( l's \) edition of Wikipedia.

3. EVALUATION

The goal of the evaluation is to illustrate the methodology of extracting the entity-centric bias of different Wikipedia language editions, and to give examples of the results and insights obtained.

3.1 Experimental Setup

To detect entity-centric language-specific bias in multilingual Wikipedia, we applied our methodology in a case study. It is well-known that the neutral point of view (NPOV) is easier to achieve in some areas, such as scientific articles, or any other uncontroversial topics.

While the pipeline presented in Section 2 is, in principle, language independent, it relies on automatic translation from the target language to English. For our experiments, which are performed using the Google Translate service, we selected Wikipedia language editions in five European languages: English (EN), Dutch (NL), German (DE), Spanish (ES) and Portuguese (PT). These editions differ in size, the largest being English (with more than 4.7 million articles), followed by German and Dutch (with about 1.8 million articles each), Spanish (about 1.1 million) and Portuguese (about 800 thousand articles).

To obtain entities that are more likely to indicate language-specific bias, we selected a total number of 219 entities with world-wide influence that come from four categories as our target entities. These four categories are: multinational corporations (55 entities), politicians (53 entities), celebrities (55 entities) and sports stars (56 entities). For each category, we included entities originating from countries that use one of the five target languages as official languages, in order to verify if the strength of the sentiments towards an entity is different in the countries of their origin.

After the Data Acquisition step described in Section 2, we created entity-centric graphs for the entities in our dataset from the five Wikipedia language editions listed above, which resulted in a total number of 1,196,403 sentences. The average number of sentences extracted from the main Wikipedia article describing the entity in our dataset is around 50. Using our data acquisition method, the number of sentences referring to the entity was increased by the factor 20 to more than 1,000 sentences per entity in a language edition. This factor is an effect of the additional data sources we use for each entity processed, as previously illustrated. Based on the Equation 2, we obtained the objective, positive and negative scores for each language edition towards the target entities.

3.2 Statistical Results

The sample set of target entities described here, and the summary of the sentiment analysis results, are presented in Table 1. Only 10 entities from each category are listed because of the space limitation.

In Table 1, “+” and “−” separately represent the average positive and negative scores of a language edition of Wikipedia towards the target entity; “count” represents the number of sentences containing the entity extracted from the specific language edition; “language” represents the official language(s) of the entity’s origin country(ies). From this table we can see that, for some entities, the number of occurrences varies a lot from language to language, but the average positive and negative scores for individual entities are in the range of [0.02, 0.09]. This indicates that, although language-specific bias exists in Wikipedia, due to the NPOV policy, this bias can be kept at a relatively low level. Moreover, controversies among different Wikipedia editions seem to be solved by allowing both positive and negative sentiment expression to co-exist, instead of removing the bias completely.

The number of occurrences of the entities in the different language editions is influenced by various factors, including the size of the Wikipedia edition, as well as the origin of the entity. Although English — the largest of Wikipedias — contains the majority of entity occurrences, some entities — like Angela Merkel, the Chancellor of Germany, and Mark Rutte, the Prime Minister of the Netherlands — are more frequently mentioned in the local Wikipedia editions. Nevertheless, we did not observe any systematic increase in the positive or negative sentences in the language corresponding to the country of the entity origin.

Our sentiment analysis results illustrate that the proportion of the objective and subjective information for any given entity in our dataset is similar across language editions and constitutes about 92%. The remaining (about 8%) contain positive and negative sentiments, which vary, depending on the particular entity and language. For example, for the named entity “Thomson Reuters”, about 6% of German Wikipedia holds positive sentiment and 3% holds negative sentiment. While in Portuguese Wikipedia, the positive sentiment score and negative sentiment score change to 4% and 3%, respectively. Maybe it is not unreasonable to say that the German-speaking people like Thomson Reuters more than the Portuguese-speaking people. For other named entities, such as “Unilever”, all the five language editions’ Wikipedia contain almost the same level of positive sentiment and negative sentiment, the score of which are 4% and 3%, respectively.

Because of the large number of named entities mentioned on Wikipedia, it is not possible to apply our approach on all of them. Based on a limited number of 219 entities, for all our five target languages, their average proportions of positive sentiment scores and negative scores for each category are at the same level. There are some other interesting patterns. For example, all the five languages average proportions of the positive and negative sentiment scores of corporations are slightly lower (about 1%) than their corresponding proportions for the people-related categories, except the average negative sentiment proportion of celebrities in Dutch Wikipedia, the average negative sentiment proportion of celebrities in German Wikipedia, the average negative senti-
Besides that, some celebrities and sports stars have relatively high positive scores. Examples are Enrique Iglesias, Taylor Swift, Shakira and Tiger Woods in the German Wikipedia. After analysing representative sentences with positive scores, we attribute this to the following reasons: First, Wikipedians are more likely to add positive sentimental terms for celebrities and sports stars, such as “Shakira’s Ojos Asi performance was chosen as the best Latin Grammy performance of all time” and “The most successful song of the year was Bailando by Enrique Iglesias”. Second, celebrities and sports stars often achieve some awards or victories, which greatly contributes to the positive sentiment scores. For example, “Tiger Woods with his 14 victories since 1997, the most successful active golfer and the second most successful in the eternal ranking” and “In addition to that, so Swift received BMI President’s Award, Which honours at exceptional individual in entertainment industry deserving of special recognition”.

In the following, we are going to analyse some of the results in more details.
3.3 Examples: GlaxoSmithKline and Angela Merkel

To analyse the aspects mentioned in the sentences with high positive and negative sentiment scores in different languages, we further explore and illustrate the automatically extracted representative sentences for two entities: GlaxoSmithKline — a British multinational pharmaceutical company, and Angela Merkel — the Chancellor of Germany.

GlaxoSmithKline occurs more frequently in the English and German Wikipedia, while less in the Dutch, Spanish and Portuguese editions. Therefore, the English and German Wikipedia mention more positive and negative aspects than the other Wikipedia editions. For example, on the one hand, many sentences from the German and English Wikipedia are about the effectiveness of the various vaccines developed by GlaxoSmithKline, but these aspects are rarely mentioned in the other Wikipedia editions. On the other hand, in the Dutch and Portuguese Wikipedia, the sentences about GlaxoSmithKline with high positive sentiment scores are mostly a rough description of the economical development of this company. This trend can also be observed within the negative aspects. While the majority of the language editions mention medicine safety issues, the company’s lawsuit and its corruption, the Portuguese Wikipedia seems quite ignorant about these incidents. Besides, the positive and negative aspects mentioned about the entity also show some level of locality. The Dutch Wikipedia mentions a local organisation Nigyo’s relationship with GlaxoSmithKline: “The organisation Nigyo who receives money from GlaxoSmithKline, has given discretion to make an EO program, which is to help patients suffering from cervical cancer” — an aspect that is mostly relevant for Dutch Wikipedia only. Similarly, the German Wikipedia mentions: “Under the umbrella of GlaxoSmithKline, Odol has become the largest oral hygiene brand in Germany” — an aspect that is mostly relevant for the German Wikipedia only.

As for Angela Merkel, the majority of the entity occurrences are located, as expected, in the German and English Wikipedia. Each of the five Wikipedia versions mention Angela Merkel’s success in the elections and some criticism she gets during her tenure. However, in the sentences with high positive scores and negative scores, different aspects have been mentioned. In the German Wikipedia, Angela Merkel receives a lot of compliments with respect to the time before she went on the political stage and became famous. For example, one sentence in the German Wikipedia describes her as "a well-known student with excellent performance in any event at the University of Leipzig". This kind of information is rarely found in other Wikipedia editions. Since the German Wikipedia includes more specific and detailed information about the facts closely related to Germany, this phenomenon is not hard to understand. Moreover, as negative sentences in the German Wikipedia, some detailed personal information can be found, regarding her haircut and clothes. English Wikipeidians seem more likely to pay attention to Angela Merkel’s relationship with other politicians, and thus include multiple comments from other politicians in the articles. For example, "I want to believe though, and I think I am right, that Angela Merkel is a fine leader with decent ethics and superior intelligence." In the Portuguese Wikipedia, a high percentage of positive sentences are the compliments to Angela Merkel’s performance in the economic crisis and on the financial market.

As these examples illustrate, the proposed methodology is effective to automatically extract entity-centric opinionated sentences from multilingual Wikipedia, as well as numerically labelling these sentences with their corresponding sentiment strength and polarisation.

4. RELATED RESEARCH

Studies on Multilingual Wikipedia: As different Wikipedia language editions evolve independently, multilingual Wikipedia became an interesting research target for different disciplines. One area of interest is the study of differences in the linguistic points of view in Wikipedia ([17] and [16]). Rogers [17] employed the multilinguality in Wikipedia as a cultural reference for a sociological case study and analysed semantic differences between the language-specific editions of the Srebrenica massacre article. This study was performed manually on one Wikipedia article only, no automation or entity tracking or extraction was attempted, as in our work.

Khatib et al. [11] detected the point of view differences between Arabic and English Wikipedia articles automatically by training different classifiers for each language. However, their method was language specific, and would require extra annotation and training to be extended to other languages, unlike our approach that has been shown to be easily generalisable.

The authors of Manypedia [11] assessed the similarity of the articles in different Wikipedia language editions, by computing their concept similarity based on the interlinking. This method operated at the article level and does not take into account occurrences of the entity in other articles, which is proposed here.

Yasseri et al. [21] performed an analysis of the most controversial articles in different Wikipedia language editions based on the edit history. This study illustrates that the most controversial entities differ in various language editions. In this work we used the list of the most controversial English entities in [21] as a starting point for our dataset generation.

Sentiment Analysis: As we determine the existence of bias based on sentiment analysis, we briefly visit here the two main processing approaches: rule-based and learning-based [11]. The rule-based sentiment analysis approach checks if the sentence complies with some predefined rules, to measure the sentiment polarity. One typical approach is to match the words occurring in the text with some sentiment lexicons, such as [8] and [20]. In this work we use SentiWordNet [2] — a free, state-of-the-art lexicon that contains more than 20,000 sentiment bearing words, due to its popularity, coverage and availability.

*http://en.wikipedia.org/wiki/Talk%3AAngela_Merkel/Archive_1
5. DISCUSSION AND CONCLUSION

In this paper we proposed a novel, easily-reproducible, automatic methodology to analyse and better understand language-specific differences in the representation of entities in different Wikipedia language editions. This methodology includes the dynamic generation of entity-centric graphs from the multilingual Wikipedia, by using and reusing in-links as well as language links and sentiment analysis techniques to better understand language-specific differences. This methodology provides some insights into the language-specific representation of individual entities.

We applied this methodology in a case study over five Wikipedia language editions (more than any predecessor), creating and analysing 219 entity graphs for the entities representing multinational corporations, politicians, celebrities and sports stars. Our results illustrate that the proportion of objective information for any given entity in our study is similar across language editions and constitutes about 92% in our dataset. The remaining 8% contain positive and negative sentiments, that vary, dependent on the particular entity and language. We observed that these 8% contained practically in all cases both positive as well as negative sentiment expressions, which may show that the neutrality in Wikipedia is obtained not by neutralising all statements, but by including a both positive and negative utterances. To better explain our results, we have further analysed some of the examples, to show that even well-known, internationally relevant entities vary quite a lot in the way they are presented in the various language editions of Wikipedia, in terms of size of the articles, focus on related aspects, and the level of sentiment attached to these entities.

In summary, we conclude that the proposed methodology is effective to automatically extract entity-centric opinionated sentences from multilingual Wikipedia, and to numerically quantify the polarity and intensity of the sentiment. For our future research, we are planning to improve the sentiment analysis techniques to be target-dependent and aspect-based, in order to get a higher accuracy. We also plan to group the Wikipedia edits not only by the languages, but also by the IP addresses, in order to achieve a finer-grained analysis of the opinion holders.

Acknowledgments

This work was partially funded by the COST Action IC1302 (KEYSTONE) and the European Research Council under ALEXANDRIA (ERC 339233).

6. REFERENCES