Is Wikipedia Really Neutral? A Sentiment Perspective Study of War-related Wikipedia Articles since 1945

Yiwei Zhou, Alexandra I. Cristea and Zachary Roberts
Department of Computer Science
University of Warwick
Coventry, United Kingdom
{Yiwei.Zhou, A.I.Cristea, Z.L.Roberts}@warwick.ac.uk

Abstract

Wikipedia is supposed to be supporting the "Neutral Point of View". Instead of accepting this statement as a fact, the current paper analyses its veracity by specifically analysing a typically controversial (negative) topic, such as war, and answering questions such as "Are there sentiment differences in how Wikipedia articles in different languages describe the same war?". This paper tackles this challenge by proposing an automatic methodology based on *article level* and *concept level* sentiment analysis on multilingual Wikipedia articles. The results obtained so far show that reasons such as people's feelings of involvement and empathy can lead to sentiment expression differences across multilingual Wikipedia on war-related topics; the more people contribute to an article on a war-related topic, the more extreme sentiment the article will express; different cultures also focus on different concepts about the same war and present different sentiments towards them. Moreover, our research provides a framework for performing different levels of sentiment analysis on multilingual texts.

1 Introduction

Wikipedia is the largest and most widely used encyclopaedia in collaborative knowledge building (Medelyan et al., 2009). Since its start in 2001, it contains more than 33 million articles in more than 200 languages, while only about 4 million articles are in English[1]. Possible sources for the content include books, journal articles, newspapers, webpages, sound recordings[2], etc. Although a "Neutral point of view" (NPOV)[3] is Wikipedia's core content policy, we believe sentiment expression is inevitable in this user-generated content. Already in (Greenstein and Zhu, 2012), researchers have raised doubt about Wikipedia's neutrality, as they pointed out that "Wikipedia achieves something akin to a NPOV across articles, but not necessarily within them". Moreover, people of different language backgrounds share different cultures and sources of information. These differences have reflected on the style of contributions (Pfeil et al., 2006) and the type of information covered (Callahan and Herring, 2011). Furthermore, Wikipedia webpages actually allow to contain opinions, as long as they come from reliable authors[4]. Due to its openness to multiple forms of contribution, the articles on Wikipedia can be viewed as a summarisation of thoughts in multiple languages about specific topics. *Automatically detecting and measuring the differences* can be crucial in many applications: public relation departments can get some useful suggestions from Wikipedia about topics close to their hearts; Wikipedia readers can get some insights about what people speaking other languages think about the same topic; Wikipedia administrators can quickly locate the Wikipedia articles that express extreme sentiment, to better apply the NPOV policy, by eliminating some edits.

[1] http://meta.wikimedia.org/wiki/List_of_Wikipedias

[2] http://en.wikipedia.org/wiki/Wikipedia:Citing_sources

[3] http://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view

[4] http://en.wikipedia.org/wiki/Wikipedia:Identifying_reliable_sources

29th Pacific Asia Conference on Language, Information and Computation pages 160 - 168
Shanghai, China, October 30 - November 1, 2015

In order to further gain insight on these matters, and especially, on the degree of neutrality on given topics presented in different languages, we explore an approach that can perform *multiple levels of sentiment analysis on multilingual Wikipedia articles*. We generate *graded sentiment analysis* results for multilingual articles, and attribute sentiment analysis to concepts, to analyse the sentiment that one*specific named entity* is involved in. For the sake of simplicity, we restrict our scenario within the war-related topics, although our approach can be easily applied on other domains. Our results show that even though the overall sentiment polarities of multilingual Wikipedia articles on the same war-related topic are consistent, the strengths of sentiment expression vary from language to language.

The remainder of the paper is structured as follows. In Section 2, we present an overview of different approaches of sentiment analysis. Section 3 describes the approach selected in this research to perform article level and concept level sentiment analysis on multilingual Wikipedia articles. In Section 4, experimental results are presented and analysed, and in Section 5, we conclude the major findings and remarks for further research.

2 Related Research

Researchers have been addressing the problem of sentiment analysis of user-generated content mainly at three levels of granularity: *sentence level* sentiment analysis, *article level* sentiment analysis and *concept level* sentiment analysis.

The most common level of sentiment analysis is the *sentence level*, which has laid the ground for the other two. Its basic assumption is that each sentence has only one target concept.

Article level (or *document level*) sentiment analysis is often used on product reviews, news and blogs, where it is believed there is only one target concept in the whole article. Our research performs article level sentiment analysis of Wikipedia articles. We believe this is applicable here, as the Wikipedia webpages' structure is that with a topic as the title, the body of the webpage is the corresponding description of the topic. There are mainly two directions for article level sentiment analysis: *analysis towards the whole article*, or *analysis towards the subjective parts* only. Through extracting the subjective sentences of one article, a classifier can not only achieve higher efficiency, because of the shorter length, but can also achieve higher accuracy, by leaving out the 'noises'. We thus choose to extract the possible subjective parts of the articles first.

More recently, many researchers have realised that multiple sentences may express sentiment about the same concept, or that one sentence may contain sentiment towards different concepts. As a result, the *concept level* (or *aspect level*) sentiment analysis has attracted more and more attention. Researchers have proposed two approaches to extract concepts. The first approach is to manually create a list of interesting concepts, before the analysis (Singh et al., 2013). The second approach is to extract candidate concepts from the object content, automatically (Mudinas et al., 2012). As in Wikipedia different articles will mention different concepts, it is impossible to pre-create the concepts list without reading all the articles. We thus choose to automatically extract the named entities in the subjective sentences as concepts.

3 Methodology

3.1 Overview

We employ war-related topics in Wikipedia as counter-examples to refute the statement that 'Wikipedia is neutral'.

Based on our choice of approaches (see Section 2), we build a straightforward processing pipeline (Figure 1) as briefly sketched below (with details in the subsequent sub-sections).

First, we retrieve the related topic name based on some input keywords. After that, the Wikipedia webpages in *all* available languages on this topic are downloaded. Because of the diversity of the content in Wikipedia webpages, some data pre-processing, as described in Section 3.2, is needed, in order to acquire plain descriptive text. We further translate the plain descriptive text into English (see notes in Section 3.2 on accuracy and the estimated errors introduced), for further processing. To extract the subjective contents from each translated article, we tokenise the article into sentences, and then perform subjective analysis, as is described in Section 3.3, on each sentence. As mentioned already, based on prior

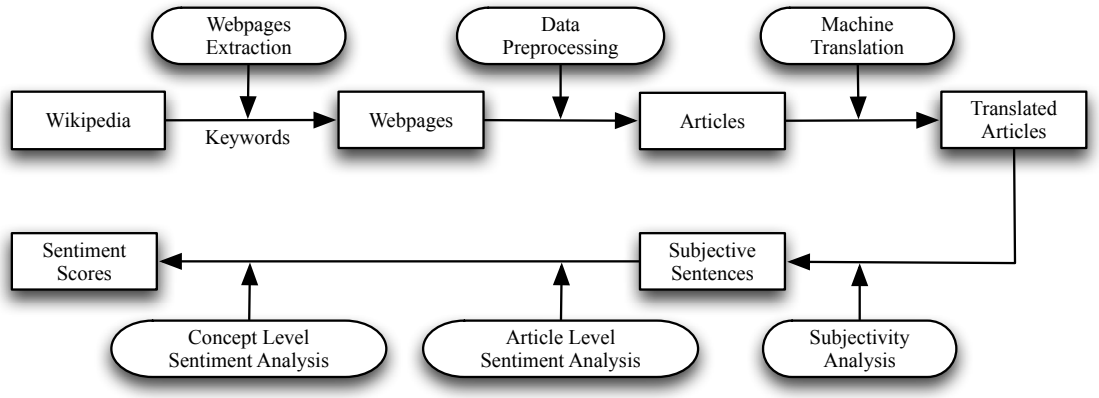

Figure 1: Processing Pipeline

research (Pang and Lee, 2004), only the subjective sentences are retained, while the rest are discarded. We then leverage the English sentiment analysis resources to measure the sentiment score for each subjective sentence, and utilise named entity extraction tools to extract the named entities as target concepts in this subjective sentence. We calculate the article level sentiment scores, as is described in Section 4.1, as well as the concept level sentence scores, as is described in Section 3.5, with both being based on the sentence level sentiment scores. In the final step, all the absolute sentiment scores of multilingual Wikipedia articles on the same topic are normalised within a range of [0, 1], for better visualisation and comparison.

3.2 Data Acquisition and Pre-processing

All the data of Wikipedia can be accessed through its official MediaWiki web API[5]. However, the downloaded webpages contain multiple kinds of information — such as references, external links, and the infobox, rather than simple plain descriptive text. Thus, the first step of data pre-processing is to discard all the parts that contain no sentiment information from the downloaded webpages . Moreover, we use the lxml[6] HTML parser, in order to remove all the HTML tags.

Whilst there are plenty of sentiment analysis tools and methods with satisfying performance for English, many of them are free, or easy to be implemented, not the same can be said for other languages. To close the gap between other languages and English sentiment analysis resources, we apply machine translation on the texts in other languages. To date, machine translation techniques are well developed; products such as Google Translate[7] and Bing Translator [8] are widely used in academic (Bautin et al., 2008; Wan, 2009) and business context. In (Balahur and Turchi, 2014), researchers pointed out that the machine translation techniques have reached a reasonable level of maturity, which could be applied in multilingual sentiment analysis. We choose Google Translate, because of its extensive use and excellent reputation. We also are able to more confidently use machine translation, after performing the following test: we translate English articles to all the other available languages for the target topic and then back to English, and we evaluate the resulting sentiment scores' changes. This test will be further discussed in Section 4.

3.3 Subjectivity Analysis

Conforming with the NPOV policy, we couldn't find clear sentiment expression in the greatest proportion of the content of the translated articles, it is the subjective parts of the content that we really care about. By extracting these parts of the content, we can compress the long Wikipedia articles into much shorter texts, with the sentiment information retained (Pang and Lee, 2004). This greatly simplifies the next pro-

[5] http://www.mediawiki.org/wiki/API:Main_page

[6] http://lxml.de/index.html

[7] https://translate.google.com/

[8] http://www.bing.com/translator/

cessing step and saved memory usage. Moreover, reliable subjectivity analysis can make the article "cleaner", by eliminating the possible errors introduced by objective sentences.

As in the next step, we will apply another more accurate tool to grade the sentiment scores of the sentences, in this stage, it is the *recall* that we focus on, rather than the *precision*.

Our proposed method thus first performs subjectivity analysis at sentence level. Since extracting as many sentences that may contain sentiment expression as possible is our first consideration, we use sentiment-bearing words' occurrences as indicators of sentiment expression in Wikipedia sentences. The detailed rule is: if one sentence contains any word from a sentiment-bearing words lexicon, then this sentence is classified as a subjective sentence; otherwise this sentence is discarded. Our method for subjectivity is based on an assumption that for a sentence, if it contains sentiment-bearing words, it may or may not be a subjective sentence; if it contains no sentiment-bearing words at all, then it is definitely not a subjective sentence. This method can greatly reduce the level of computational complexity and maintain a high recall.

Liu et al. (Liu et al., 2005) created a list of positive and negative opinion words for English, which has about 6800 words, including most of the adjectives, adverbs and verbs that contain sentiment information used on the web. This list fully satisfies all our needs, thus is used in our subjectivity analysis.

3.4 Article Level Sentiment Analysis

It is the overall sentiment score of each Wikipedia article rather than the separate sentiment scores of sentences in the article that we care about in this research. For example, if *Article A* has 10 positive sentences and 2 negative sentences, and *Article B* has 5 positive sentences and 3 negative sentences, we assume that *Article A* is more positive than *Article B*, thus will have a higher absolute sentiment score. It should be noted that we do not normalise the article level sentiment scores by the numbers of sentences here because the numbers of positive/negative sentences can also reflect the sentiment levels of different articles. Similar to (Ku et al., 2006) and (Zhang et al., 2009), we calculate the sentiment score $S(a)$ of an article a, by aggregating the sentiment scores of the subjective sentences in it, as follows:

$$S(a) = \sum_{i=1}^{m} S(s_i, a) \tag{1}$$

where: $S(s_i, a)$ denotes the sentiment score of the i^{th} subjective sentence s_i in article a, and m denotes the number of subjective sentences in article a.

There are a lot of applicable sentence level sentiment analysis tools, and it is essential to choose the one most suits for this context. As different sentences will express different levels of sentiment, it is better to use the sentiment analysis tool that can estimate such a difference, rather than only classifies the sentences as positive or negative. Moreover, the chosen sentiment analysis tool should have acceptable performance in measuring the sentiment levels of sentences on war-related topics.

The Stanford CoreNLP sentiment annotator in Stanford natural language processing toolkit(Manning et al., 2014) can reach an accuracy of 80.7% on movie reviews (Socher et al., 2013). The annotator will classify the sentiment of each sentence into five classes, including very negative, negative, neutral, positive and very positive. To verify its performance on war-related sentences, we generate a list of English sentences randomly selected from war-related Wikipedia articles. After manually labelling 200 sentences, we use Stanford CoreNLP sentiment annotator to generate graded results for our labelled data. Its accuracy on war-related sentences is 72%, which satisfies our needs for this application. For calculation convenience, we assign each class a sentiment score from -2 to 2, where -2 represents very negative, -1 represents negative, 0 represents neutral, 1 represents positive and 2 represents very positive. The sentiment scores of sentences will be aggregated later into the overall sentiment score of each translated article according to Equation 1. In short, if S_a is greater than 0, it means this is a positive article; if S_a is equal to 0, it means this is a neutral article; otherwise this is a negative article. Besides that, the scores also give account of the levels of sentiment involved.

3.5 Concept Level Sentiment Analysis

After computing the sentiment scores of articles on the same topic in multiple languages, we expand our

research to finer granularity, the concept level sentiment analysis. By exploring what concepts are mentioned and their corresponding sentiment scores in one article, we expect to locate the underlying reasons of why articles show different levels of sentiment. Inspired by (Mudinas et al., 2012), we extract the named entities from the subjective sentences, as the concepts mentioned in the article. In (Atdag and Labatut, 2013), Atdag and Labatut compared different named entity recognition tools' performance, and the Stanford NER[9] showed the best results. Thus we apply Stanford NER to extract the concepts. We use the sentences that one concept occurs in as its opinionated context (as in (Singh et al., 2013; Mudinas et al., 2012)).

The sentiment score $S(c, a)$ of each concept c in article a is calculated as follows:

$$S(c, a) = \sum_{j=1}^{n} S(s_j, c, a) \quad (2)$$

where: $S(s_j, c, a)$ denotes the sentiment score of the j^{th} subjective sentence s_j in article a which mentions the concept c, and n denotes the number of subjective sentences in article a which mentions the concept c. As mentioned in Section 4.1, if $S(c, a)$ is greater than 0, it means concept c is more involved in positive sentiment; if $S(c, a)$ is equal to 0, it means concept c has a overall neutral context; otherwise the concept is more involved in a negative sentiment. Similar to article level sentiment score, we do not apply normalisation here either because the number of positive/negative sentences that mention this concept can also reflect the level of sentiment this concept involved in.

4 Evaluation

We choose war-related topics as a start for the neutrality analysis of multilingual Wikipedia for the following reasons. First, Wikipedians have different sentiment expression patterns for topics from different domains. While it is not possible to perform the multilingual Wikipedia sentiment differences analysis for all these domains, we choose one domain as a start. If the the NPOV of Wikipedia cannot hold for this domain, by providing these topics as counter-examples , Wikipedia is not neutral in general (although there exist many neutral articles). Second, war-related topics are controversial in the first place. For different belligerents of the wars, they often use different official languages, and have different interpretations towards the same incidents, which makes the detection of sentiment differences possible. Third, as illustrated in Section 4.1, Stanford CoreNLP sentiment annotator has acceptable performance on the sentences from the domain of wars, but its performance on the sentences from other domains remains unknown.

To analyse sentiment differences in the perception of war-related topics, we compare sentiment scores of multilingual Wikipedia articles, perform concept level sentiment analysis and explore the relationship between the sentiment scores and numbers of words/concepts in the articles, as described below.

4.1 Article Level Sentiment Differences in Multilingual Wikipedia Articles on War-related Topics

We have performed article level analysis on all the wars with clear belligerents and a certain level of popularity since the ending of the *Second World War*. There are 30 of them satisfy our demands from the list of wars provide by Wikipedia[10]. Due to page limitation, the results of 7 of them can be found in Table 1.

There are 666 Wikipedia pages in 68 languages on these 30 war-related topics, 100% of them have an overall negative sentiment. This shows consistency of sentiment polarity of multilingual Wikipedia articles on war-related topics. In Table 1, a ranked list is given, starting from the most neutral language to the most negative language, thus the languages in the first half of the ranked list have articles more neutral than the languages in the second half of the ranked list on a specific war-related topic; the official languages of belligerents are marked in *italic* characters.

To measure the influence of Google Translate on the final results, we design one test: for each one the 30 war-related topics, we translate its English edi-

[9] http://nlp.stanford.edu/software/CRF-NER.shtml

[10] http://en.wikipedia.org/wiki/Category:Lists_of_wars_by_date

Table 1: Sentiment Differences in Multilingual Wikipedia on War-related topics.

War-related topics	**Languages' ranked list (from most neutral to most negative)**
Korean War	Japanese, Nepali, Hindi, Afrikaans, Malay, Macedonian, Esperanto, Armenian, Tamil, Welsh, Bengali, Swahili, Belarusian, Azerbaijani, Basque, Persian, Latin, Serbian, Arabic, Hungarian, Greek, Romanian, Norwegian, Turkish, Lithuanian, Slovak, Filipino, Icelandic, Thai, Danish, Bosnian, Croatian, Estonian, Galician, Dutch, Latvian, Polish, Swedish, Czech, Spanish, Mongolian, Finnish, Bulgarian, Ukrainian, Slovenian, Portuguese, German, Indonesian, Telugu, Kannada, *Russian*, Italian, French, Vietnamese, *Korean*, *Chinese*, *English*
Algerian War	Greek, Romanian, Malay, Bengali, Persian, Esperanto, Irish, Basque, Portuguese, Spanish, Swedish, Vietnamese, Welsh, *Arabic*, Lithuanian, Korean, Chinese, Croatian, Catalan, Turkish, Polish, Hungarian, Serbian, Norwegian, Dutch, Finnish, Japanese, Czech, Latvian, Russian, Italian, Ukrainian, German, *French*, English
Turkish invasion of Cyprus	Serbian, Arabic, Polish, Czech, Romanian, Norwegian, Persian, Korean, German, Chinese, Portuguese, Hungarian, French, Spanish, Russian, *Turkish*, Swedish, Finnish, Italian, *Greek*, English
Dirty War	Esperanto, Finnish, Swedish, Tamil, Korean, Welsh, Ukrainian, Georgian, Polish, Russian, Malay, Portuguese, Bosnian, Persian, Chinese, Japanese, Czech, Serbian, Croatian, Italian, Indonesian, German, Dutch, French, Galician, *Spanish*, English
Romanian Revolution of 1989	Basque, Indonesian, Irish, Croatian, Arabic, Thai, Norwegian, Czech, Japanese, Swedish, Slovak, Korean, Chinese, Russian, Turkish, Serbian, Portuguese, French, Ukrainian, Bulgarian, Filipino, Finnish, Dutch, Polish, Catalan, Spanish, Galician, Italian, Hungarian, English, *Romanian*, German
Civil war in Tajikistan	Bosnian, Italian, Portuguese, Serbian, Norwegian, Catalan, Bulgarian, Welsh, Polish, German, Ukrainian, Spanish, Japanese, French, English, Czech, *Russian*
War in North-West Pakistan	Korean, *Urdu*, Czech, Portuguese, Spanish, Croatian, Welsh, Hungarian, German, Russian, Japanese, French, *English*, Polish

tion Wikipedia article to all its other available languages on Wikipedia, then we translate the translated articles back to English. Then we calculate the sentiment scores of these translated-then-back articles and get the new ranks of them. After this process, we find that all these 30 war-related topics satisfy: if the English edition Wikipedia article is in the first/second half of the ranked list, all its translated-then-back articles remains in the first/second half of the ranked list. This test shows Google Translate's impact on our final result is quite limited.

People from the belligerents usually suffer the most from the wars, so we expect the official languages of belligerents have the most negative sentiment towards the wars. Our results show: of these war-related topics we tested, 80% share a common characteristic, which is the official languages of the belligerents have relatively more negative sentiment towards the wars (rank in the second half of the ranked list) than other non-relevant languages. The results we get are quite consistent with our expectations, which also prove the effectiveness of our method. For example, Russian is one of the belligerents of Civil war in Tajikistan; it has the most negative sentiment towards this topic. US, China and Korea are greatly involved in the Korean War, and the corresponding Wikipedia editions are most negative on this topic. This is the same as in the case of French on the Algerian War, Greek on the Turkish invasion of Cyprus, Spanish on the Dirty War and Romanian on the Romanian Revolution of 1989. On the contrary, the most neutral Wikipedia articles about Civil war in Tajikistan are in Bosnian, Italian and Portuguese. This may be caused by the following reasons: first, Bosnian Wikipedia is not as widely used as some other languages (ranked 69 in number of Wikipedia articles[11]), which will result in a limited number of edits on these articles; second, the Civil war in Tajikistan has little relevance to the people speaking, e.g., Italian or Portuguese, because of the geographical distance, which may result in limited attention to this topic. Our findings, besides corroborating our method, also present some other interesting patterns that are worth exploring. For example, on the topic of Korean War, we can also see that Vietnamese, a language used among only 75 millions people[12], also holds very negative senti-

[11] http://meta.wikimedia.org/wiki/List_of_Wikipedias

[12] http://en.wikipedia.org/wiki/

ment. This may due to the geographical distance between Vietnam and Korea, but also because the Korean War is very close in time to the Vietnam War. Thus, the extreme sentiment of Vietnamese people towards the Korean War may not be surprising at all. On the topic of the Romanian Revolution of 1989, German and Hungarian Wikipedia hold very negative sentiment. This is because during the period of the Romanian Revolution of 1989, there are also similar revolutions in Hungary and East Germany. Some kind of empathy makes German and Hungarian Wikipedia users have similar sentiment as the Romanian Wikipedia users. On the topic of War in North-West Pakistan, Polish pages have the most negative sentiment. We can speculate that Polish people feel involved in this war, as a Polish engineer was kidnapped and killed by Pakistani extremists.

However, some official languages of belligerents seems to be not that negative towards the wars they were involved in. For example, the sentiment of Arabic Wikipedia on the Algerian War and the sentiment of Urdu Wikipedia on the War in North-West Pakistan. The possible reasons can be summarised as follows. First, according to the statistics provided by Wikipedia [13], the Arabic Wikipedia and Urdu Wikipedia is far less active than some other languages, such as English and German. Second, these languages' sentiment expression patterns may be largely different from English, thus the sentiment analysis resources for English may not work well on these languages' translated articles. A detailed analysis of sentiment expression patterns on linguistic level is beyond the scope of this work.

4.2 Concept Level Sentiment Analysis on English Wikipedia and Russian Wikipedia on the Civil war in Tajikistan

Table 2 lists the concepts extracted from French Wikipedia on Civil war of Tajikistan. For comparison, Table 3 lists the concepts extracted from the Russian Wikipedia on the same topic. To add readability, we only keep the concepts with absolute sentiment scores no less than 2. The method of calculating the sentiment scores of concepts can be found in Section 3.5.

From Table 2 and Table 3, concepts that are involved in negative sentiment in both French Wikipedia and Russian Wikipedia on the topic of Civil war in Tajikistan are marked in *italic* characters. Obviously there are more concepts involved in negative sentiment in Russian Wikipedia than French Wikipedia, which is understandable since Russian is the the most widely used language among the countries that are involved in the war. People from these Russian speaking countries have more detailed information about the war, and rich sentiment towards the war, thus will mention more concepts in its corresponding Wikipedia article and express stronger sentiment in the contexts of these concepts.

There are some concepts that occur only in French Wikipedia, but not in Russian Wikipedia. For example, Abdullo Nazarov, Rasht Vally and Movement for Islamic Revival of Tajikistan. Similarly, a lot more concepts occur only in the Russian Wikipedia but not in the French Wikipedia. Concepts occurring only in specific languages editions of Wikipedia may point to people's variances in preference and focus.

Table 4: Pearson correlation coefficient between sentiment scores and articles' features

Results	p_1	p_2
Average	0.971	0.948
Standard deviation	0.020	0.025

4.3 Relationship Between Sentiment Scores and Number of Words/Concepts in Multilingual Wikipedia Articles

To analyse the underlying reasons leading to the differences in sentiment level, we calculate the Pearson correlation coefficient between the article level sentiment scores and some features of the articles. The article features we choose are the number of words in the article and the number of concepts mentioned in subjective sentences in the article. The statistical summary of results of 30 war-related topics we test is displayed in Table 4. In the table, p_1 is the Pearson correlation coefficient between sentiment scores and numbers of words; p_2 is the Pearson correlation coefficient between sentiment scores and numbers of concepts.

Vietnamese_language

[13] http://meta.wikimedia.org/wiki/List_of_Wikipedias

Table 2: Concept Level Sentiment Analysis of French Wikipedia about the Civil war in Tajikistan.

Concept Type	**Concepts involved in Negative Sentiment**
Person	Abdullo Nazarov, Tolib Ayombekov, *Mullah Abdullah*
Location	*Afghanistan*, *Dushanbe*, *Garmi*, *Gorno-Badakhshan*, Rasht Valley, Samsolid, *Tajikistan*, Taloqan, *Uzbekistan*
Organisation	Movement for Islamic Revival of Tajikistan, Taliban, *United Tajik Opposition*

Table 3: Concept Level Sentiment Analysis of Russian Wikipedia about the Civil war in Tajikistan.

Concept Type	**Concepts involved in Negative Sentiment**
Person	Dawlat Khudonazarov, Emomali Rahmon, Karim Yuldashev, Mahmoud Khudayberdiev, Mirzo Ziyoyev, Mukhiddin Olimpur, *Mullah Abdullah*, Nozim Vahidov, Otakhon Latifi, Rahmon Nabiyev, Rahmon Sanginova, Safarali Kenjayev, Saifullo Rakhimov, Victor Khudyakov, Yusuf Iskhaki
Location	*Afghanistan*, Darwaz, *Dushanbe*, *Garmi*, *Gorno-Badakhshan*, Hissar, Iran, Karategin, Kazakhstan, Khujand, Kofarnihon, Kulyab, Kulob Oblast, Kurgan-Tube, Kyrgyzstan, Leninabad, Lomonosov, Majlisi Oli, Nurek Dam, Ozodi, Pakistan, Shakhidon, *Tajikistan*, Tavildara, *Uzbekistan*, Vakhsh
Organisation	Afghan Mukahideen, CIS, Communist Party, Democratic Party of Tajikistan, Islamic Renaissance Party, Lali Badakhshan, Rastokhez, National Guard, Tajikistan Interior Ministry, *United Tajik Opposition*

The Pearson correlation coefficient measures the strength of linear correlation between the articles' features and the sentiment scores of these articles from different Wikipedia editions. A Pearson correlation coefficient of nearly 1 means there is strong positive correlation between the two variables. 100% of the Pearson correlation coefficients between sentiment scores and numbers of words of multilingual Wikipedia articles (p_1), and 96.7% of the Pearson correlation coefficients between sentiment scores and numbers of named entities of multilingual Wikipedia articles (p_2) are above 0.9. This illustrates that for the war-related articles in Wikipedia, the more words in one negative article, the more negative the article will be; the more concepts in subjective sentences in one negative article, the more negative the article will be. Both the number of words and the number of concepts in one translated article reflect the degree of concern of people speaking that language about this topic. A higher degree of concern will drive people to add more contents to the Wikipedia article about that topic in their language, which will lead to stronger sentiment expression in corresponding article.

5 Conclusion

Is Wikipedia really neutral? By using war-related topics as proof by counter-examples, we find the short answer to this question: no.

Our results demonstrate that, while multilingual Wikipedia articles on one war-related topic have a consistent sentiment polarity, there are differences on levels of sentiment expression. People's degree of concern about one war-related topic will influence the number of words, and the number of subjective concepts, which in turn determine the levels of sentiment expression. The subjective concepts mentioned and their frequencies also reflect the fact that people speaking different languages have different focuses and interests about the same war-related topic. For some languages, there is no obvious connection between them and the belligerent countries at first glance; nevertheless, they often have more extreme sentiment towards the war than other irrelevant languages. When discrepancies happen, some underlying reasons can always be found by thoroughly researching into the war history. Since it is not possible to ask people to read all the Wikipedia articles in different languages on the same topic and rank them based on their sentiment expression levels, we validate our results through qualitative analysis, by locating the underlying reasons that lead to such results.

While our findings only apply to war-related topics on Wikipedia, our approaches can be further applied on various topics and domains to explore the sentiment differences of multilingual Wikipedia.

References

Samet Atdag and Vincent Labatut. 2013. A comparison of named entity recognition tools applied to biographical texts. In *Systems and Computer Science (ICSCS), 2013 2nd International Conference on*, pages 228–233. IEEE.

Alexandra Balahur and Marco Turchi. 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1):56–75.

Mikhail Bautin, Lohit Vijayarenu, and Steven Skiena. 2008. International sentiment analysis for news and blogs. In *ICWSM*.

Ewa S. Callahan and Susan C. Herring. 2011. Cultural bias in wikipedia content on famous persons. *Journal of the American Society for Information Science and Technology*, 62(10):1899–1915.

Shane Greenstein and Feng Zhu. 2012. Collective intelligence and neutral point of view: The case of wikipedia.

Lun-Wei Ku, Yu-Ting Liang, and Hsin-Hsi Chen. 2006. Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 100107.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 342–351. ACM.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Olena Medelyan, David Milne, Catherine Legg, and Ian H. Witten. 2009. Mining meaning from wikipedia. *International Journal of Human-Computer Studies*, 67(9):716–754.

Andrius Mudinas, Dell Zhang, and Mark Levene. 2012. Combining lexicon and learning based approaches for concept-level sentiment analysis. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM '12, pages 5:1–5:8. ACM.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04. Association for Computational Linguistics.

Ulrike Pfeil, Panayiotis Zaphiris, and Chee Siang Ang. 2006. Cultural differences in collaborative authoring of wikipedia. *Journal of Computer-Mediated Communication*, 12(1):88–113.

V.K. Singh, R. Piryani, A Uddin, and P. Waila. 2013. Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. In *2013 International Multi-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)*, pages 712–717.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 235–243. Association for Computational Linguistics.

Changli Zhang, Daniel Zeng, Jiexun Li, Fei-Yue Wang, and Wanli Zuo. 2009. Sentiment analysis of chinese documents: From sentence to document level. *Journal of the American Society for Information Science and Technology*, 60(12):2474–2487.