KNOWLEDGE – International Journal
Vol.31.6
June, 2019

LANGUAGE CHALLENGES IN ASPECT-BASED SENTIMENT ANALYSIS: A REVIEW OF ALBANIAN LANGUAGE

Majlinda Axhiu
South East European University, Republic of North Macedonia, majlinda.axhiu@gmail.com

Abstract: Beside the advantages of the typical sentiment analysis, which focuses on predicting the positive or negative polarity of the given sentence(s), there are two main drawbacks of performing sentiment analysis on higher level, namely on sentence and document level. Firstly, gaining the overall sentiment of a sentence or a paragraph may not lead to accurate and precise information. The polarity will be valid for a broader context and not for particular targets. Secondly, many sentences or paragraphs may have opposing polarities towards different targets. This makes it difficult or impossible to give an accurate overall polarity.
The necessity for detecting aspect terms and their corresponding polarity gave rise to aspect-based sentiment analysis (ABSA). To meet the objectives of aspect-based sentiment analysis systems, the process can be summarized in three main tasks: Aspect Term Extraction, Aspect-term and Opinion-word Separation and Sentiment Polarity Classification.
Most commonly, supervised learning approaches are used for ABSA. However, having to build the tagged training and testing corpora for each language and each domain is highly time consuming and can often be achieved only manually. This is why we have used a semi-supervised model for designing a language- and domain-independent system that is based on novel machine learning approaches through which we are focused on analyzing Albanian texts and make use of Albanian data in the digital world.
In this approach where we try to extract the aspects and the polarity of their corresponding opinions through almost unsupervised learning, the biggest challenge is to reach high accuracy in natural language processing. In order to achieve this, in language-independent systems there must be taken into consideration all the differences and similarities of the languages.
In this paper our aim is to define the biggest challenges that appear in Albanian language in comparison with English; and after analyzing certain amount of data, we have identified the following issues: inflections, negation, homonyms, dialects, irony, sarcasm and stop-words' presence in aspect terms. This is not an exhaustive list of the language issues, since we have selected and discussed only the ones that have greater impact in the process of extracting the aspect-terms and opinions, and can highly affect the accuracy of the final polarity classification of the texts.
Keywords: ABSA system, aspect-term extraction, language challenges, language-independent systems

1. INTRODUCTION

Most of the published sentiment analysis results (on websites or reports) only show the aggregated ratings of products, but people are more interested in detailed opinions that capture aspect specific features in reviews/comments. Therefore, it is desirable to have an all-inclusive approach to mine aspect specific opinion expressions containing both aspect terms and opinion words within the sentence context as a composite aspect-sentiment, and additionally group them under logical aspect categories. Except from finding out the key advantages and disadvantages of the products or services that are often expressed via aspect-sentiment phrases, they are also useful in applications such as comparing similar products and summarizing their important features where it is more convenient to have the aspect-sentiment phrases rather than generic aspect/sentiment words lacking the natural aspect opinion correspondence in the right context (Laddha and Mukherjee, 2016).
Supervised aspect term extraction using human annotated datasets reaches high performance for aspect term detection on previously unseen data. However, this method has two major disadvantages. First, human annotation is a very costly and slow process, and second, the size of the labeled datasets is relatively small. The size of the datasets may consequently reduce the performance of the classifiers. These disadvantages may be surpassed by using unsupervised aspect term extraction. Regarding the first disadvantage, the usage of an automated data labeling process with very high precision can substitute the costly and slow human annotation process. In order to decrease incorrectly annotated aspect terms in the automated labeling only nouns and noun phrases are considered (Giannakopoulos, Musat, Hossmann & Baeriswyl, 2017).
On the other hand, taking into consideration the size of the used datasets, it can be expressively increased by using publicly available data/reviews. Taking into account that reviews or comments from the online world are

opinionated texts and contain many different aspects, they become good candidates for building new datasets for aspect term extraction.

The texts we have analyzed in the scope of the paper are from social media, namely from Facebook. In order to have data from multiple domains we have collected Facebook comments from one of the biggest Albanian portals in Macedonia, Tetova1's fan page. After cleansing the data, the final number of comments on which we have worked on is 31.250 comments.

2. LANGUAGE CHALLENGES

Although both Albanian and English languages are part of the Indo-European family, taking into account the phonetic, semantic and grammatical system, they have a lot of differences.

We have identified several language challenges that are important to be mentioned and taken into consideration in the tasks of aspect-based sentiment analysis, such as: inflections, negation, homonyms, dialects, irony, sarcasm and stop-words' presence in aspects.

2.1. INFLECTIONS AND WORD ORDER

The first issue that we have recognized is the fact that Albanian language in contrast to English is highly inflectional language. The high number of various word forms makes Albanian a hard language to deal with. Previous studies state that for highly inflectional languages like Albanian, stemming or lemmatization is almost obligatory because it is essential to decrease the high number of diverse word forms.

Another great challenge for Albanian language is the flexible word order. In contrast to this, English has a very fixed word order, and if in any case the word order is changed, the sentence would change its meaning as well, and probably will become non or less acceptable or not in a good grammatical condition.

2.2. NEGATION

Negation is another issue which differs in English and Albanian languages. English uses single negation, whereas Albanian language besides single negation, it may use double or even triple negations. For instance,

- "**S'**vjen **asnjeri**" - (double negation)
- "**Askush askujt s'i** ka borxh" – (triple negation)
- "**Asnjëherë asgjë nuk** i kam thënë." – (triple negation)

This factor makes it difficult for machines to understand the real meaning of the sentences. Another differentiation in terms of negation is that in English we add auxiliaries, while in Albanian we are not.

2.3. HOMONYMS

The next issue that we came up within our data was the presence of homonyms. There are three problematic cases regarding the homonyms. First, when they are nouns (e.g. "bari" which has three meanings: shepherd, grass and medicament), in the aspect term extraction process they will be extracted as one term, which definitely will lead to wrong results.

Secondly, when one of the homonyms is noun and the other verb (e.g. "vesh" –meaning "ear" or "put on" (some clothes)), the noun should be considered and extracted as an aspect terms, while the verb should be not, because it isn't a term. However, there is a very high probability that the system will consider both of them as aspect terms and will aggregate them.

The third case is when one of the homonyms is an aspect term while the other one is a stop word (e.g. "dhe"- meaning "soil" or "and"). In the first meaning, when the word appears as a noun it should be extracted as an aspect term, while in the case when it is used as a conjunction it should be neglected. Here the risk is that these cases wouldn't be considered at all as aspect terms, just because they will be skipped, since they are part of stop words.

2.4. DIALECTS AND REGIONAL SPEECH VARIETIES

Having two main dialects and tens of regional speech varieties, Albanian language becomes really challenging for the tasks of aspect-based sentiment analysis. The chances are really high that some of the aspects (which may be with high frequency) may not be extracted and considered due to the different spelling or even completely different words' usage (e.g. qumësht-tomël, rrugë-udhë, atëror-atnor, vatër-votër, grua-grue, etc.).

A specific example that we have found in our data, which changes the complete context of the text is due to the regional speech varieties. In some regions in Macedonia the letter "y" is replaced with "i", so for instance in our case the word "dy" (meaning "two") is written as "di" which in formal language means "know". Due to this confusion, the whole sentence is neglected, which results in losing of meaningful data.

Considering all these issues that may appear from the usage of dialects and/or regional speech varieties, there have been conducted specific researches on multi-dialect sentiment analysis systems, which result to have higher accuracy than the others and also greater coverage of the data for that specific language.

2.5. IRONY AND SARCASM

Irony and sarcasm are special forms of speech in which the opinion holders write the opposite of what they mean. While irony is often used to emphasize happenings that deviate from the expected, sarcasm is commonly used to carry implied criticism. However, the recognition of irony and sarcasm is a difficult task, even for humans. The trouble in recognizing irony and sarcasm causes misunderstanding in everyday communication and states problems to many natural language processing tasks such as sentiment analysis. This is especially challenging when we deal with social media messages, where the language is short and informal.

Since even at their definition it is stated that it is written the opposite meaning, the probability is very high that also the sentiment polarity may be opposite to the real meaning. For example,

- "Të gjithë duken normal, deri sa i njofton!" (In English: "Everyone looks normal, until you meet (them)!"

Here according to the structure and the initial meaning the statement is positive, however the real meaning is negative (none is normal).

2.6. SENTENCES WITH AMBIGUOUS MEANING

There exist a lot of ambiguous sentences that may have different sentiment polarities in different cases (mainly depending from the opinion holder). For instance "I love children" may have a positive or negative sentiment depending if the opinion holder is a child abuser or not; similarly the sentence "I love sweets, and I can eat at every opportunity!" again may have positive or negative sentiment, depending from the author if he/she is diabetic or not.

There is another category of sentences which appear to be affirmative according to the structure, but negative in meaning. For instance,

- "Ku dihet?"= "Nuk dihet",
- "Ku di unë!" = S'di unë.",
- "E kush pyet për mua?" ="Asnjë s'pyet për mua."

The same issue, where the polarity of the sentiment is difficult to be detected is also the opposite case, where the sentence seems to be negative according to the structure, but it has a positive meaning. For example: "E ç'nuk kishte aty!"= " Kishte nga të gjitha aty!"

The last detected ambiguous sentence has to do with the usage of some particles such as: "sikur", "mbase" and "ndoshta", which neutralizes the negative/positive meaning of the sentences. For instance:

- "Sikur nuk ndjehem mirë"
- "Mbase nuk vij as unë."

2.7. STOP-WORDS WITHIN ASPECT TERMS

Even though the aim of the usage of STOP-words is to increase the accuracy of the results, namely in aspect term extraction, there are some exceptional cases when they can have the opposite effect.

In the section of homonyms, we have already mentioned that some stop words may appear the same as some aspects, however since the focus on that point were the single-word aspects; we can assure that those cases are very rare.

The issue and the challenges become greater when we have to do with multiple-word aspect terms, where the presence of the stop-words is significant. For instance in the cases of "këpucë për fëmijë", "sapun për tesha", "baltë e kuqe", etc., where the words "për" and "e "are already in the list of stop-words, their consideration can be destructive because the structure of aspect terms will be divided and most probably they will be considered as two different aspect terms (e.g. këpucë and fëmijë, sapun and tesha).

3. CONCLUSION

In the scope of the paper there are discussed the issues that can affect the precision of aspect-terms extraction and the extraction of opinions for each aspect, depending on the language. We analyzed an Albanian corpus and pointed out a lot of challenges in the aspect of language structure, grammar and semantics, such as: homonyms, multi-word aspects, inflections, ironic expressions, etc.

Taking into account that both Albanian and English language are coming from the same family and still having a lot of differences, we should be more aware for other languages that may be included in multilingual sentiment analysis systems, because all of these differences among languages can affect the result of the ABSA tasks, even in language independent systems.

BIBLIOGRAPHY

Biba, M., Mane, M. (2015). Sentiment Analysis through Machine Learning: An Experimental Evaluation for Albanian. Volume 235 of the series Advances in Intelligent Systems and Computing, pg. 195-203

Giannakopoulos, A., Musat, C., Hossmann, A., & Baeriswyl, M. (2017). Unsupervised Aspect Term Extraction with B-LSTM and CRF using Automatically Labelled Datasets. Proceedings Of The 8Th Workshop On Computational Approaches To Subjectivity, Sentiment And Social Media Analysis. doi: 10.18653/v1/w17-5224

Laddha, A., & Mukherjee, A. (2016). Extracting Aspect Specific Opinion Expressions. In Empirical Methods in Natural Language Processing (pp. 627–637). Austin, Texas: Association for Computational Linguistics.

Pablos, A., Cuadros, M., & Rigau, G. (2017). W2VLDA: Almost Unsupervised System for Aspect Based Sentiment Analysis. Arxiv.org. Retrieved from https://arxiv.org/abs/1705.07687