Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), pages 159–166
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Component Analysis of Adjectives in Luxembourgish for Detecting Sentiments

Joshgun Sirajzade[1], Daniela Gierschek[2], Christoph Schommer[1]
[1]MINE Lab, Computer Science, Université du Luxembourg
[2]Institute of Luxembourgish Linguistics and Literatures, Université du Luxembourg
{joshgun.sirajzade, daniela.gierschek, christoph.schommer}@uni.lu

Abstract

The aim of this paper is to investigate the role of Luxembourgish adjectives in expressing sentiments in user comments written at the web presence of rtl.lu (RTL is the abbreviation for Radio Television Lëtzebuerg). Alongside many textual features or representations, adjectives could be used in order to detect sentiment, even on a sentence or comment level. In fact, they are also by themselves one of the best ways to describe a sentiment, despite the fact that other word classes such as nouns, verbs, adverbs or conjunctions can also be utilized for this purpose. The empirical part of this study focuses on a list of adjectives which were extracted from an annotated corpus. The corpus contains the part of speech tags of individual words and sentiment annotation on the adjective, sentence and comment level. Suffixes of Luxembourgish adjectives like *-esch*, *-eg*, *-lech*, *-al*, *-el*, *-iv*, *-ent*, *-los*, *-bar* and the prefix *on-* were explicitly investigated, especially by paying attention to their role in regards to building a model by applying classical machine learning techniques. We also considered the interaction of adjectives with other grammatical means, especially other part of speeches, e.g. negations, which can completely reverse the meaning, thus the sentiment of an utterance.

Keywords: Opinion Mining / Sentiment Analisys, Corpus (Creation, Annotaton, etc.), Grammar and Syntax

1. Introduction

Detecting the sentiment of an utterance has been dealt with in numerous publications and using different machine learning techniques. A lot of the tools were built for languages with a large number of speakers such as English, French and German. For smaller languages like Luxembourgish, well-trained and established tools are still rather scarce. We utilize a large Luxembourgish corpus and extract a subset that we annotated for sentiment on comment, sentence and adjective level. The aim of the paper is to leverage this data source in order to explore feature combinations other than semantic similarity representations for detecting sentiment, primarily by analyzing adjectives and their components. As Luxembourgish is a low-resource language, no resources for sentiment detection have been built so far. Intuitively, adjectives carry a high amount of sentiment. Examining them and their components in our data subset could therefore provide important insights into how much they could potentially help to improve our system's performance. This paper first gives an overview over existing resources and research in sentiment analysis and over the Luxembourgish language. We then portray our annotation process and dig deeper into our corpus to look at the adjectives and their suffixes that we subsequently use for our experiments. To conclude, we discuss future work that could further help research on the importance of adjectives for sentiment analysis.

1.1. Research and Resources in Sentiment Analysis

Sentiment analysis has been seen for a long time as a pure text classification problem (Pang et al., 2002) whereas recent research in the area has brought light to many details and other forms of it. Placing it as a part of mining of opinions and emotions, it was shown that sentiment analysis can have different levels, e.g. sentence level vs. aspect based sentiment analysis (Liu, 2015). While classification, especially deep classification, still gives the best results, there are attempts to customize the text classification problem to the needs of sentiment analysis, e.g. by creating sentiment specific word embeddings (Tang et al., 2014). Word embeddings, alongside other bag of words techniques for text representation like Latent Semantic Analysis utilize the so called distributional similarity, in other words they calculate the semantic similarity of words based on their distribution in text data (Levy et al., 2015). This approach touches the semantics from a linguistic point of view, yet the usage of other levels of language are still to be investigated. Other methods for detecting sentiment include lexical approaches that use manually or automatically constructed dictionaries containing positive and negative words and sometimes even more granular description of sentiment, e.g. the strength of the polarity. Those dictionaries are then used to calculate the overall sentiment of the unseen data (Taboada, 2016). An example for a lexical resource is SentiWordNet (Esuli and Sebastiani, 2006) which was built for English and assigns either positive, negative or objective to synsets[1] of WordNet. For Luxembourgish, no resource of this kind exists yet which brings special difficulties to be tackled for implementing such an approach for this language.

Despite a large number of publications dealing with sentiment analysis and its different aspects, challenges that need to be solved still remain. Attempts have been made to use automatically translated data (Balahur and Turchi, 2012) or to (semi-)automatically create a sentiment corpus querying Twitter data for certain emojis (Pak and Paroubek, 2010). Very often however, sentiment detection systems are based on the manual labeling task of one or more annotators. Those annotators need to be recruited, trained and provided with adequate guidelines which makes this part of the

[1]Synsets are unordered sets of synonym words that denote the same concept and are interchangeable in many contexts, see https://wordnet.princeton.edu.

system construction time and resource intensive. Creating labels for sentiment very much depends on the guidelines given, as it is not as simple as just giving a positive, negative or neutral score to an entity and not always easy for an annotator to stay consistent in his/her annotation. Therefore, clear and simple instructions are crucial for ensuring the best annotation possible (Mohammad, 2016). One big challenge is that words can have very different meanings depending on their context (Mohammad, 2016). If we look at the adjective *stolz* [proud] for example, it conveys a very different sense in those two contexts: *Ech sinn stolz drop.* [I am proud of that.] vs. *Do bass du stolz drop???* [You are proud of that???]. These two different meanings would probably be impossible to catch in a lexical approach where an annotator would annotate *stolz* isolated from its context. Also, a sentiment can be directed towards the reader, the speaker or the writer of an utterance (Mohammad, 2016). It therefore has to be clearly stated in the annotation guidelines how the annotation is supposed to be undertaken. In this paper, we focus on the role of adjectives in sentiment analysis as they carry a lot of the subjective aspects of a text (Taboada, 2016) and thus bear a high sentiment content. More precisely, we focus not only on adjectives but especially on some specific suffixes and one prefix of adjectives in Luxembourgish and how those might have an impact on detecting the sentiment of sentences.

1.2. Luxembourgish Language

Luxembourgish is mainly spoken in the Grand Duchy of Luxembourg, a multilingual country with roughly 590,000 inhabitants (Gilles, in press). Despite Luxembourg having three official languages, i.e. French, German and Luxembourgish, only the latter was recognized as the unique national language of the country in 1984 and has become an important symbol for national identity (Gilles, in press) since. It developed out of a Central Franconian dialect and is thus related to German. However, Luxembourgish today is perceived as an independent language by the speech community (Gilles, 2015). The language plays an important role in spoken and written conversation. If all participants of a discussion are capable of using this language, Luxembourgish can be used in any formal or informal situation and code-switching to another language would be unimaginable (Gilles, in press). The importance of fostering the Luxembourgish language can be seen in several projects across the country. One prominent example is the *Schnëssen* app which was developed at the Institute of Luxembourgish Linguistics and Literatures to preserve the current varieties and ways of speaking Luxembourgish. Crowdsourcing techniques are leveraged for recording as many spoken examples of Luxembourgish as possible. Those are then used to portray the speakers' variation on different linguistic levels (Entringer et al., 2018).

On an NLP level, the LuNa Open Toolbox (Sirajzade and Schommer, 2019) was implemented as a rule-based part-of-speech tagger and tokenizer especially designed for dealing with Luxembourgish texts and their linguistic characteristics. LuNa is essential for working with Luxembourgish texts as it is the only tool so far that was built for processing and dealing with the special challenges related to this low-ressource language. Texts in digital media, such as user comments that we will investigate in this paper or text messages, are almost exclusively produced in Luxembourgish. This is remarkable, as the educational system mainly focuses on German and French and not that much on the orthographic rules of Luxembourgish (Gilles, 2015). Not focusing on Luxembourgish spelling in school results in high orthographic variation in texts such as our corpus data. Variation in spelling is a great challenge for our project. A lot of written data in Luxembourgish exists, but a big part of it is not spelled according to the official spelling rules. Using a lexical approach for Luxembourgish sentiment analysis would therefore be very labor-intensive because no tool that captures every kind of possible variation of Luxembourgish spelling has been built yet.

2. Annotation of the Data Source

For our project, we have obtained the database of rtl.lu (Radio Télévision Lëtzebuerg), a popular news website that mostly publishes in Luxembourgish (RTL Luxembourg, 2019). It consists of over 180,000 news articles from 1999 to 2018 and over 500,000 user comments from 2008 to 2018. More precisely, our corpus comprises more than 30 million running tokens for the news articles part and over 35 million running tokens for the comments part.

2.1. Corpus Creation

In a first step, we tokenized our whole corpus (Sirajzade and Schommer, 2019) and also undertook part-of-speech tagging and sentence splitting. We then used part of our database and asked one annotator to annotate this subcorpus on document (comment), sentence and word (adjective) level. The guideline was to annotate from the perspective of the author (Abdul-Mageed and Diab, 2011) and to use the labels *positive, negative* and *neutral*. Comments were randomly chosen to ensure that the training corpus would not just consist of sentiments towards a single topic. The sentences and adjectives in those comments were then also tagged with their sentiment value. Our data is stored in *XML*. During the annotation process, two new tags were introduced: <*comment*> and <*sentence*>. The annotator also included an attribute *value* into those two new tags and for the adjectives. Furthermore, she provided the attribute with its respective sentiment value, i. e. *positive, negative* or *neutral*. Figure 1 shows an example of an annotated user comment in our corpus in *XML*. Considering that we only had one annotator, no inter-annotator agreement was calculated. The dataset we used for our analysis is discussed in more detail in the following section.

2.2. Subcorpus

The annotated subcorpus that we use for our investigation in this paper is composed of 431 comments, 2050 sentences and 1339 adjectives. 132 comments were marked as positive, 208 as negative and 91 as neutral. On sentence level, the annotator perceived 499 as positive, 833 as negative and 718 as neutral. 574 adjectives of the ones annotated were tagged with a positive, 327 with a negative and 438 with a neutral value. Our special focus for this analysis lies on the annotated adjectives that we extracted from the

```
<comment value ="positive">
    ...
    <sentence value = "neutral">
        <w id="36" pos="P">Hien</w>
        <w id="37" pos="AUX">huet</w>
        <w id="38" pos="AV">do</w>
        <w id="39" pos="D">e</w>
        <w id="40" pos="N">grouse</w>
        <w id="41" pos="ADJ" value="neutral">perséinleche</w>
        <w id="42" pos="N">Konflikt</w>
        <c id="43" pos="$">,</c>
        <w id="44" pos="APPR">op</w>
        <w id="45" pos="P">hien</w>
        <w id="46" pos="D">sengem</w>
        <w id="47" pos="N">Gewëssen</w>
        <w id="48" pos="KO">oder</w>
        <w id="49" pos="D">senger</w>
        <w id="50" pos="N">Flicht</w>
        <w id="51" pos="KO">als</w>
        <w id="52" pos="N">Staatschef</w>
        <w id="53" pos="V">follegt</w>
        <c id="54" pos="$">.</c>
    </sentence>
</comment>
```

Figure 1: Example of an annotated user comment from the RTL corpus in *XML*

corpus for further fine-grained analysis. Using 1339 adjectives is likely too small to draw conclusions for all adjectives present in the Luxembourgish language. However, we expect to gain a first intuition concerning the impact of leveraging adjectives as features. Figure 1 shows an example of the extracted data from our corpus, which later serve as data instances for our machine learning experiments. It is important to note that the adjectives were annotated in their context and therefore do not always carry the same sentiment for all times they were annotated. For instance, *typesch* [typical] (see figure 2) was annotated three times whereas it was negative in two and positive in one case.

3. Distribution of Grammatical Properties

As a first step of setting up our experiment, we looked at the adjectives and counted all suffixes and prefixes that were annotated in our data. We then also investigated the occurrence of negation in the corpus. This information is essential for getting a better understanding of our data for the experiments we undertook.

3.1. The Distribution of Adjectives in Sentiments

Even though certain adjectives can utter different sentiments depending on the pragmatic or syntactic context, it can be observed that several adjectives have a tendency towards a certain sentiment. Tables 2 and 3 show the ten most frequent adjectives once by their own sentiment and once by the sentiment of the sentences they were used in. The ambiguity of adjectives in expressing the sentiment of the sentences becomes especially clear in the words *richteg*

Luxembourgish	English	Sentiment	Sentence Sentiment	Comment Sentence
sarkastesch	sarcastic	negative	negative	neutral
chinesesch	Chinese	neutral	neutral	neutral
europäesch	European	neutral	neutral	neutral
historesch	historical	neutral	negative	neutral
praktesch	practical	positive	negative	positive
komesch	strange	negative	negative	neutral
demokratesch	democratic	positive	negative	positive
pornografesch	pornographic	negative	neutral	neutral
gigantesch	gigantic	negative	negative	negative
typesch	typical	negative	positive	negative

Figure 2: Luxembourgish adjectives with the suffix *-esch* and their sentiments & sentiments of the respective sentence and comment

[right], *besser* [better] and *einfach* [simple]. They occur almost the same amount of time in both the positive and the negative categories. This phenomenon stretches, as mentioned before, from the pragmatic level, where the sentiment of the adjective in itself can be ambiguous depending on the intention of the author (which can be for example sarcastic) up to the syntactic structure of the utterance through combination with negation, which can instantly change the sentiment. The next step is to investigate the internal structure of the adjectives – their suffixes and prefixes and to look if using them as features can lead to some important generalization.

3.2. Suffixes and Prefixes

For this paper, we extracted some of the most important suffixes and one prefix of the adjectives in our corpus in order to study their importance for sentiment detection. More precisely, we used the suffixes *-esch*, *-eg*, *-lech*, *-al*, *-ent*, *-el*, *-iv*, *-los*, *-bar* and the prefix *on-* for this analysis. We chose those five suffixes and one prefix, because they are prominent for word formation processes in Luxembourgish language (Sirajzade, 2018). However, we need to keep in mind that word formation elements are generally not that frequently distributed and out of the 1339 adjectives in our corpus only 289 have one of these elements (see table 4). We will use those adjectives for our analysis. Besides looking at them individually, we also noticed what the sentiment of the sentence and comment they were found in was. In table 5 the relationship between the suffix of a particular adjective and the sentiment of the sentence are shown. The suffix *-esch* is mostly present in neutral adjectives such as *komesch* [funny], but builds rather negative sentences and comments. More positive adjectives are assembled using the suffix *-eg*, like for example in *spaasseg* [amusing]. As stated in section 3.3., those adjectives usually occur in negative sentences and comments which can be seen as a strong indication of the importance of negation in those contexts. The suffix *-lech*, like in *ënnerschiddlech* [different], mostly appears in neutral adjectives whereas sentences or comments are often positive or negative. *-bar* is a suffix which seems to have a strong tendency towards positivity. Adjectives like *tragbar* [portable] were mostly annotated as positive or neutral and so were sentences and comments that were almost exclusively perceived as positive. The last suffix we examined, *-los* like in *skrupellos* [unscrupulous], has a tendency towards negativity. We only found very few positive sentences or comments that included an adjective with this suffix. Most adjectives, sentences and comments were annotated as negative. The prefix *on-* that we examined is mostly used for negative adjectives such as *onfair* [unfair] and also negative sentences and comments. This is not surprising however as *un-* in itself reverses the meaning of an adjective to some extend (see section 3.3.).

3.3. Negation

There is only little evidence of negation in our small subcorpus that we have created for this experiment. Out of the 1339 sentences which contain adjectives, only 337 appear in the context of some sort of negation. We considered the negation particle *net*[not] and indefinite pronoun *kee(n)*[no]

	Adjective	Suffix	Negation	Adjective's Sentiment	Sentence's Sentiment
1	richteg [right]	-eg	/	positive	positive
2	flexibel [flexible]	-el	/	positive	neutral
3	héich [high]	/	net	positive	negative
4	illegal [illegal]	-al	/	negative	negative
5	diktatoresch [dictatorial]	-esch	net	negative	negative
6	eenzel [single]	-el	net	neutral	negative
7	domm [stupid]	/	net	negative	negative
8	perséinlech [personal]	-lech	/	neutral	negative
9	anonym [anonymous]	/	/	neutral	positive
10	éierlech [honest]	-lech	/	positive	positive

Table 1: The features adjective, its suffix, its sentiment and the sentiment of the sentence it is used in in the experiments

positive	Frequency	negative	Frequency	neutral	Frequency
besser [better]	23	schlecht [bad]	18	laang [long]	23
gudd [good]	23	falsch [wrong]	12	perséinlech [personal]	7
richteg [right]	21	deier [expensive]	9	kleng [small]	6
einfach [easy]	15	lues [slow]	7	grouss [big]	5
gutt [good]	13	traureg [sad]	7	groussen [big]	5
wichteg [important]	12	laang [long]	6	lang [long]	5
grouss [big]	9	blöd [stupid]	5	krank [sick]	4
kloer [clear]	8	domm [stupid]	5	normal [normal]	4
genau [exact]	7	egal [same]	4	nächst [next]	4
gudden [good]	7	komesch [strange]	4	nächsten [next]	4

Table 2: The 10 most frequent positive, negative and neutral adjectives

in our analysis. Nevertheless, we could make two interesting observations that should be examined further in future experiments. First of all, adjectives with the suffix *-eg* like in *wichteg* [important] were mostly annotated as positive, but very often occur in negative sentences. Negation thus seems to play an important role for expressing negativity in combination with an adjective that carries the suffix *-eg*. Adjectives with the prefix *on-* like in *onwichteg* [unimportant] were mostly annotated as negative and did not occur with negation in a sentence. This is interesting as *on-* already carries negativity and can reverse the sentiment of an adjective to negative. For instance, ommitting the prefix *on-* from the adjective "onwichteg" [unimportant] would result in the positive adjective "wichteg" [important]. It is therefore not surprising that we did not find any kind of double negation in sentences with *on-* adjectives.

4. Experiment

After having looked into our data, we explore different supervised machine learning settings using a combination of different features. The goal of the experiments will be to examine the role of adjectives and their suffixes in the overall sentiment of a sentence it appears in.

4.1. Setup of the Experiment

We build different kinds of scenarios in order to investigate the role of adjectives in the building of the sentiment of a sentence. We have a total of four features and one label, which we combine in different ways. The features (more precisely feature groups) are the adjective (ADJ), its suffix (SUFF), negation in the sentence (NEG), and the adjective's sentiment (ADJ-SEN). The label is the sentiment of the sentence in which it is used (SENT-SEN). We decided not to include prefixes as a feature, as we only found one, i.e. *on-*, in our data and do not consider this sufficient for representing Luxembourgish prefixes in general. This structure is shown with the first ten instances of the data in table 1. We created one-hot vectors from ADJs, SUFFs and NEGs, so each adjective, suffix or negation particle is a feature in itself. Note that we did not use the TF-IDF vectorizer (except for comparison purposes in the eleventh scenario) or any other similar technique for this particular experiment, because the setup assumes that only certain part of speeches e.g. the adjectives and negation within the sentences are known. Because of the fact that a sentence in our dataset contains in average one and only seldomly two or more adjectives, we do not count their occurrences. Furthermore, we assume that by just seeing one adjective, it is possible to determine the sentiment of the sentence, so every adjective is considered as its own instance. For our experiments, we used the *scikit-learn (0.21.2)* library in *Python* (Pedregosa et al., 2011) and *WEKA (3.8.2)* (Hall et al., 2009). Both environments have implementations of many commonly known machine learning algorithms which can be applied in sentiment analysis. With some differences, which was the reason why we experimented with both of them, they are very suitable for testing purposes. We used 10-fold cross validation and optimized the gamma and the c value for SVM by using the Radial Basis Function (RBF) kernel.

positive	Frequency	negative	Frequency	neutral	Frequency
gudd [good]	12	laang [long]	14	grouss [big]	7
einfach [easy]	10	schlecht [bad]	13	laang [long]	7
richteg [right]	10	richteg [right]	10	wichteg [important]	7
besser [better]	9	besser [better]	9	besser [better]	5
laang [long]	9	einfach [easy]	9	gudd [good]	5
groussen [big]	5	gudd [good]	7	lang [long]	5
gutt [good]	5	lues [slow]	7	falsch [wrong]	4
kleng [small]	5	deier [expensive]	6	groussen [big]	4
wichteg [important]	5	grouss [big]	6	gutt [good]	4
falsch [wrong]	4	spéit [late]	6	kloer [clear]	4

Table 3: The 10 most frequent positive, negative and neutral adjectives by their sentence sentiment

The results of the experiments are presented in table 6. We carried out experiments in 11 different scenarios. For each scenario we used Decision Tree (DT), Support Vector Machine (SVM) and Complement Naive Bayes (CNB) from *scikit-learn* and Bayes Net (BN) from *WEKA*. We included DT in our experiment, because our data with the adjectives has a more categorical or nominal character and it is easy to interpret. SVM has been a standard algorithm for sentiment analysis for a long time. It is very suitable and effective in a high dimensional space which we have after vectorizing our data. We experimented with CNB and BN because we additionally wanted to test a predictive model which in the case of BN can also capture Markov states throughout our features. To examine the performance of each algorithm, we then calculated the weighted average of precision, recall and F1 score. Weighted F1 score is calculated as

$$\frac{1}{\sum_{l \in L} |\hat{y}_l|} \sum_{l \in L} |\hat{y}_l| \, F(y_l, \hat{y}_l),$$

where y_l is the subset of predicted labels, $\hat{y}$ the subset of true labels and L the set of labels. Precision and recall are then calculated in the same manner as described in (scikit-learn, 2019). Table 6 uses the commonly known machine learning notation *X* for describing the features used for a scenario and *y* for the label that was to be classified. The first scenario is the simplest one; there are only one feature (group) and one label. Those are the adjective and the sentiment of the sentences in which it is used. More precisely, this scenario classifies the sentiment of a sentence only by looking at the adjective that is contained in this sentence. Note that it is not the sentiment of the adjective itself, but of the sentence in which it was used. This approach was already tried on many languages, but was never implemented for the Luxembourgish language before. The weak point of this method is the fact that despite the big role that adjectives play in the building of sentiment of the utterance, there are some other language elements which can change or reverse the sentiment of the adjective. Those are, for instance, grammatical negation or the intentional use of sarcasm. For this reason, we first look at the role of the adjective in determining the sentiment of a sentence. Then we consequently add supplementary information, beginning with the suffix, the sentiment of the adjective itself and the negation. Those scenarios are shown in rows one to five in table 6. In the third scenario, we look for example at the adjective and its own sentiment as feature. This is directly connected to the semantics of the adjective and often different than the sentiment of the sentences it is used in. As mentioned before, the sentiment of words can be easily changed by some grammatical and pragmatical means. This change of sentiment was important for the annotation process described in 2.2. In the second, fourth and fifth scenarios we use the above mentioned suffixes as an additional feature, if the respective adjective contains one. Beginning in the sixth scenario, we try to determine if suffixes of adjectives play a role in the building of sentiments

	Suffix	Frequency	positive	negative	neutral
1	-eg	117	61	33	23
2	-lech	89	37	15	37
3	-esch	61	8	20	33
4	-al	36	9	11	16
5	-el	17	8	3	6
6	-ent	14	6	6	2
7	-iv	14	11	0	3
8	-bar	6	2	1	3
9	-los	3	0	3	0
	Prefix				
1	on-	22	0	19	3

Table 4: Suffixes, prefixes and their frequency according to the sentiment of the adjective

	Suffix	Frequency	positive	negative	neutral
1	-eg	117	35	55	27
2	-lech	89	27	40	22
3	-esch	61	11	32	18
4	-al	36	14	15	7
5	-el	17	4	10	3
6	-ent	14	4	7	3
7	-iv	14	6	4	4
8	-bar	6	4	1	1
9	-los	3	1	2	0
	Prefix				
1	on-	22	0	16	6

Table 5: Suffixes, prefixes and their frequency according to the sentiment of the sentence

	Scenario	Algorithm	Precision	Recall	F1 score
1	X = ADJ y = SENT-SEN	DT	0.617	0.403	0.465
		SVM	1.0	0.421	0.592
		CNB	0.533	0.353	0.382
		BN	NaN	0.427	NaN
2	X = ADJ, SUFF y = SENT-SEN	DT	0.538	0.374	0.425
		SVM	1.0	0.421	0.592
		CNB	0.444	0.363	0.374
		BN	NaN	0.427	NaN
3	X = ADJ, ADJ-SEN y = SENT-SEN	DT	0.51	0.511	0.51
		SVM	1.0	0.421	0.592
		CNB	0.509	0.511	0.509
		BN	0.552	0.508	0.506
4	X = ADJ, SUFF, ADJ-SEN y = SENT-SEN	DT	0.501	0.5	0.5
		SVM	1.0	0.421	0.592
		CNB	0.512	0.514	0.513
		BN	0.552	0.508	0.506
5	X = ADJ, SUFF, NEG y = SENT-SEN	DT	0.671	0.41	0.484
		SVM	1.0	0.421	0.592
		CNB	0.375	0.317	0.324
		BN	NaN	0.427	NaN
6	X = SUFF y = SENT-SEN	DT	1.0	0.421	0.592
		SVM	1.0	0.421	0.592
		CNB	0.672	0.342	0.397
		BN	NaN	0.427	NaN
7	X = SUFF, NEG y = SENT-SEN	DT	0.978	0.424	0.585
		SVM	1.0	0.421	0.592
		CNB	0.49	0.317	0.348
		BN	NaN	0.427	NaN
8	X = SUFF, ADJ-SEN y = SENT-SEN	DT	0.52	0.514	0.512
		SVM	0.541	0.529	0.527
		CNB	0.541	0.529	0.527
		BN	0.552	0.508	0.506
9	X = SUFF, NEG, ADJ-SEN y = SENT-SEN	DT	0.515	0.511	0.512
		SVM	0.541	0.529	0.527
		CNB	0.532	0.522	0.519
		BN	0.552	0.508	0.506
10	X = ADJ, SUFF, NEG, ADJ-SEN y = SENT-SEN	DT	0.509	0.5	0.503
		SVM	1.0	0.421	0.592
		CNB	0.502	0.504	0.502
		BN	0.552	0.508	0.506
11	X = tf-idf vectors from sentences y = SENT-SEN	DT	0.436	0.376	0.39
		SVM	1.0	0.362	0.531
		CNB	0.631	0.388	0.445
		BN	0.371	0.397	0.301

Table 6: Different scenarios of determining the sentiment with the help of adjectives with different algorithms

of sentences. Our tenth scenario uses all the available features, the adjective, its suffix, negation and the sentiment of the adjective to determine the sentiment of the sentence. The last scenario is a special one. For this we indeed created separately tf-idf vectors from the training text without considering any part of speeches. This is a classical way of how a text classification would be done and should serve as a comparison baseline.

4.2. Results of the Experiment

Table 6 shows the precision, recall and F1 score we achieved for our experimental setups using *scikit-learn* or *WEKA*. When looking at the results produced by BN, a couple of NaNs can be seen. Those signify that the corresponding value could not be calculated due to a denominator of 0. There is no best performing result for all scenarios, different algorithms react differently to the change of features. Interestingly, the first thing to notice is that all the scenarios perform similar or better than the 11th scenario with tf-idf vectorization. The reason for this is the small size of the

data. SVM and BN seem to be the most resistant against change in the features. CNB is the algorithm that profits the most from the additional number of features. The most important observation however lies in the fact that the results in the sixth up to ninth scenarios do not drop substantially although the number of features are drastically reduced by removing adjectives and replacing them with suffixes from 1388 to around 15 depending on the scenario (10 endings, 2 negation particles, 3 adjective sentiments). Using adjectives and their suffixes together has mostly a negative influence on DT and CBN, because they both contain basically the same information, with suffixes being artificially withdrawn from the adjectives but presenting the information in a more general way. The tenth setup uses all possible features, i.e. the adjective, its sentiment, its suffix and the negation for determining the sentiment of a sentence. All algorithms perform relatively well in this scenario. An additional interesting point in the results lies in the fact that using a lexical approach indeed gives better results than using tf-idf values when dealing with a small amount of data. This could be useful especially in the case of low-resource languages.

To sum up, our results show that leveraging suffixes as an additional feature does not necessarily improve the performance of the classification system. Comparing the scenarios after the sixth to the previous ones demonstrates though that suffixes as features can replace adjectives while the algorithms give similar and comparable results. Especially DT delivers a good performance when using suffixes as an only feature to classify the sentiment of a sentence. Replacing adjectives with its suffixes results in a huge feature reduction, which is easy to maintain and can be very useful in the case of a low-resource language. However, the amount of annotated adjectives, as seen in 3.2., is rather small. In future work, we will have to annotate more data to explore whether or not the amount and diversity of suffixes available has an impact on the performance of our system or not.

5. Future Work and Outlook

We showed the importance of word formation elements for detecting the sentiment of adjectives in this paper. They can supply the same or similar amount of information as the adjectives themselves. The same should be done for other word classes, especially for the ones with more complex morphology like verbs and nouns. Using morphological information could give the same performance as using words without a need for a large annotated corpus. It is in a way a generalization which could be used for unseen words. That is why we propose a hybrid system for the Luxembourgish language which works combining language rules and machine learning techniques. However, we only worked on a relatively small sample. In the near future, we will annotate more word classes and include more suffixes and prefixes to investigate whether this can improve the performance of our system even further. We plan to integrate more annotators with the help of crowdsourcing and investigate the inter-annotator agreement. As we only annotated the prefix *on-* in our data, we will focus particularly on including a variety of prefixes (Luxembourgish verbs, for instance, have more of them.) for further experiments. Additionally, we would like to compare it to a deep learning version of our experiments in order to investigate whether or not that kind of approach can lead to promising results. Similar approaches have been implemented for deep learning, e.g. fastText (Bojanowski et al., 2016), which can also include sub-word information using n-grams. Nevertheless, in this technology the information is again repeated by creating the n-grams. Additionally, these are still not hybrid approaches and do not use linguistic rules, but rather try to learn it from the data, which could be insufficient in the case of a low-resource language. So far we have used various feature and label combinations. When working with a low-resource language such as Luxembourgish, it is important to not forget to plan enough time for studying its syntax and for the annotation process. Despite it being time consuming, we believe that it is better than translating already existing resources from other languages, as e.g. adjectives can carry different sentiments in different cultures and languages. As described in section 1.2., Luxembourgish texts usually contain lots of spelling variation, which is also very typical for low-resource languages. When dealing with this kind of data, an intensive preprocessing step could be useful.

6. Acknowledgements

This study is part of *STRIPS - A Semantic Search Toolbox for the Retrieve of Similar Patterns in Luxembourgish Documents*, an interdisciplinary project between the MINE Lab and the Institute of Luxembourgish Linguistics and Literatures at the University of Luxembourg. The aim of the project is to develop a toolbox of semantic search algorithms for texts written in Luxembourgish with a special focus on detecting sentiment (Gilles et al., 2019). The project combines machine learning techniques with linguistic knowledge for its work.

7. Bibliographical References

Abdul-Mageed, M. and Diab, M. (2011). Linguistically-motivated subjectivity and sentiment annotation and tagging of Modern Standard Arabic. *International Journal on Social Media MMM: Monitoring, Measurement, and Mining*.

Balahur, A. and Turchi, M. (2012). Multilingual sentiment analysis using machine translation? In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 52–60, 07.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*.

Entringer, N., Gilles, P., Martin, S., and Purschke, C. (2018). Schnëssen-App - Är Sprooch fir d'Fuerschung.

Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC*, volume 6, pages 417–422. Citeseer.

Gilles, P., Schommer, C., Sirajzade, J., Purschke, C., and Gierschek, D. (2019). STRIPS - A Semantic Search Toolbox for the Retrieve of Similar Patterns in Luxembourgish Documents.

Gilles, P. (2015). From status to corpus: Codification and implementation of spelling norms in Luxembourgish. *In W., Davies and E., Ziegler (Eds.), Macro and micro language planning (pp. 128-149). London: Palgrave Macmillan (2015).*

Gilles, P. (in press). Luxembourgish. *In P., Maitz, H. C., Boas (Ed.), A., Deumert (Ed.) and M., Louden (Ed.), Varieties of German Worldwide. Oxford: Oxford University Press.*

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 12.

Liu, B. (2015). *Opinions, Sentiment, and Emotion in Text*. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge University Press.

Mohammad, S. M. (2016). Challenges in Sentiment Analysis. In *A Practical Guide to Sentiment Analysis*. Springer.

Pak, A. and Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of LREC*, volume 10, 01.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

RTL Luxembourg. (2019). https://www.rtl.lu/. Accessed: 2020-01-20.

scikit-learn. (2019). https://scikit-learn.org/. Accessed: 2020-01-22.

Sirajzade, J. and Schommer, C. (2019). The LuNa Open Toolbox for the Luxembourgish Language. *In Petra Perner (Ed.), 19th Industrial Conference, ICDM 2019 New York, USA, July 17 to July 21 2019, Poster Proceedings 2019, Advances in Data Mining, Applications and Theoretical Aspects*.

Sirajzade, J. (2018). Korpusbasierte Untersuchung der Wortbildungsaffixe im Luxemburgischen. Technische Herausforderungen und linguistische Analyse am Beispiel der Produktivität. *Zeitschrift für Wortbildung = Journal of Word Formation*, 1:195–216.

Taboada, M. (2016). Sentiment Analysis: An Overview from Linguistics. *Annual Review of Linguistics*, 2(1):325–347.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., and Qin, B. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1555–1565.