Languages’ Impact on Emotional Classification Methods

Alexander C. Eilertsen, Dennis Højbjerg Rose, Peter Langballe Erichsen, Rasmus Engesgaard Christensen, Rudra Pratap Deb Nath

Department of Computer Science, Aalborg University, Selma Lagerlöfs Vej 300, 9220 Aalborg Ø, Denmark
Email: {aeiler16, drose16, perich16, rech16}@student.aau.dk & rudra@cs.aau.dk

Abstract—There is currently a lack of research concerning whether Emotional Classification (EC) research on a language is applicable to other languages. If this is the case then we can greatly reduce the amount of research needed for different languages. Therefore, we propose a framework to answer the following null hypothesis: The change in classification accuracy for Emotional Classification caused by changing a single preprocessor or classifier is independent of the target language within a significance level of $p = 0.05$. We test this hypothesis using an English and a Danish data set, and the classification algorithms: Support-Vector Machine, Naive Bayes, and Random Forest. From our statistical test, we got a $p$-value of 0.12852 and could therefore not reject our hypothesis. Thus, our hypothesis could still be true. More research is therefore needed within the field of cross-language EC in order to benefit EC for different languages.

Keywords: Sentiment Analysis, Emotional Classification, Text-to-Emotion Analysis, Cross-Language Analysis, Natural Language Processing

I. INTRODUCTION

The research field of Sentiment Analysis (SA) focuses on textual analysis, concerning the underlying emotions behind language [1]. Emotional information is extracted by using a variety of different methods. This can be used for a number of purposes, e.g. opinion mining during elections.

SA contains the subfield: Emotional Classification (EC), which focuses on classifying the emotions expressed through a medium. For EC, we use the base emotions defined by [2]: joy, trust, fear, surprise, sadness, disgust, anger, and anticipation.

A vast majority of the SA research uses English as the target language. However, it is currently not known whether the results of this research also are applicable to other languages (i.e. cross-language applicability). If the results of SA research based on one target language are applicable to SA for other languages, then that will be very beneficial for SA on non-English languages. We define this area as cross-language EC. To the best of our knowledge no one has conducted research within this area.

Based on this we specify the following null hypothesis: The change in classification accuracy for Emotional Classification caused by changing a single preprocessor or classifier is independent of the target language within a significance level of $p = 0.05$. We test this hypothesis, through an experiment that utilizes a framework we create. This framework consists of three overall phases: Preprocessing phase, Classification Phase (CP), and Statistical Test Phase (STP). The preprocessing phase consists of three subphases: Common Preprocessing Phase (CPP), Varying Preprocessing Phase (VPP), and Attribute Selection Phase (ASP). This framework serves as a guide for researchers to create experiments with similar structure and purpose as the one we are doing in this study. We do this experiment in order to test whether the effectiveness of different EC methods, trained using tweets, depend on the language being classified.

This experiment uses two data sets; one for Danish and one for English. These data sets consist of posts from the microblogging website Twitter.com, called ‘tweets’. Tweets are reasonable EC data candidates because they have the purpose of sharing emotions. They are often labeled with keywords, called hashtags, which can be emotional words such as ‘happy’. Furthermore, tweets have a character limitation of 280 characters, which entails a higher density of emotions per word.

We compare the differences in impact of changing preprocessors and classifiers on the two data sets, by applying these differences on a two-sided Wilcoxon signed-rank test, from now on referred to as ‘Wilcoxon test’.

The result from the Wilcoxon test yields a $p$-value of 0.12852, which does not reject our hypothesis. Therefore, it is still possible that EC research on the English language, is applicable to EC on non-English languages. However, since this is only a single experiment, with one non-English language, then more cross-language research is necessary to determine this.

The remainder of the paper is structured in the following way: In Section II we look into previous research within the field of EC. Section III then clarifies the definitions used in this study. Our framework as well as our application of it is defined in Section IV. Details of our experiment are then specified in Section V. In Section VI we present and evaluate our experiment results. The consequences and potential error sources of our results are discussed in Section VII. The conclusion of our study as well as ideas for further research are shown in Section VIII.
II. RELATED WORKS

During this Section, we introduce a list of EC studies, with a different focus compared to us. We also explain which elements of these studies we use for our experiment.

The main difference between these studies and ours is that while most of these sources examined different preprocessing methods and classification algorithms for the English language, we are comparing preprocessing methods and classification algorithms across multiple languages, in order to check the impact the languages have on their effectiveness.

[3] studied which preprocessing technique yields the highest accuracy using a Naïve Bayes Multinomial (NBM) classifier. They used a set of common preprocessors (i.e. preprocessors used in all test cases), and varying preprocessors (i.e. preprocessors which varied whether they were used or not). The combination that yielded the best result, when classifying positive and negative sentences, was the set of common preprocessors and stemming. Using this setup, they were able to achieve an accuracy of 80%.

In [4] they compared accuracies of multiple different n-gram combinations as well as other features, including preprocessing methods and various lexical resources. Their experiment used LIBLINEAR and NBM as classification algorithms. Based on their research we decided to test the following n-gram combinations: $NG = \{1\}, NG = \{1, 2\}$, and $NG = \{1, 2, 3\}$.

[5] presented a method for classification using anger, disgust, fear, joy, sadness, and surprise as base emotions, as well as classifying positive, negative, and neutral emotions. The classification was done using a Support-Vector Machine (SVM) classifier with Sequential Minimal Optimization (SMO) calculated on a cluster of computers, and yielded results with accuracies between 65% and 85% depending on the preprocessing methods used. We decided to use some of the preprocessing methods described in [5].

The effectiveness of different SA classification algorithms using tweets was studied by [6]. Based on their research we chose to use Random Forest (RF) and SVM as our classification algorithms. We chose these since we wanted classifiers which performed well and with very different behaviors to cover a wide spectrum of classifiers. RF was overall stable and gave good results, and is chosen as a reliable classifier, whereas SVM showed high performance as a binary classifier, but was shown to be highly data set dependent on 3 class classification.

A framework for detecting emotions in multilingual text was presented by [7]. They developed their emotion extraction system from features that were acquired from different emotional lexicons. Emotions were classified on data gathered from real-time events in different domains, such as sports.

Based on the before mentioned research we chose to use five of the preprocessing methods from [5] and two of the classification algorithms from [6]. We also chose to work with Naïve Bayes (NB) as it is a common classification algorithm. We also used the n-gram preprocessing method with the n-gram combinations that performed best in [4]: $NG = \{1\}$, $NG = \{1, 2\}$, and $NG = \{1, 2, 3\}$. While there are many studies on EC for a single language, there is a lack of research on cross-language EC. The main focus of our research is to address this issue.

III. PRELIMINARY DEFINITIONS

The definitions we need to clarify are:

- Cross-language: Applying research based on one language to other languages.
- Attribute: Unique word/n-gram from our data set.
- Instance: A tweet from our data set.
- Class: A base emotion from: \{joy, trust, fear, surprise, sadness, disgust, anger, anticipation\}[2].
- VPP configuration: A specific combination of preprocessing methods, used in VPP.
- Classification configuration: A combination of a VPP configuration and a classifier.
- Test case configuration: A combination of a classification configuration and a target language.
- Test case: An instance of a test case configuration, including the data set and the results of classifying this data set.

IV. OUR PROPOSED FRAMEWORK

In this Section, we define the framework for the general point of view as well as how we apply the framework to our experiment.

A. Framework

The framework is designed to classify a number of test cases. Afterwards, we use a statistical test on these results to evaluate whether the languages used in the data sets have a significant impact on the preprocessors and classifiers being tested.

The input of the framework is a customizable set of data sets in different languages, preprocessing methods, and classification algorithms. Preprocessing methods are divided into common preprocessors and variable preprocessors. Common preprocessors are applied to all test cases, while variable preprocessors are tested as part of the experiment.

We define the framework by three phases: Preprocessing phase, Classification Phase (CP), and Statical Test Phase (STP). The preprocessing phase consists of the following subphases: Common Preprocessing Phase (CPP), Varying Preprocessing Phase (VPP), Attribute Selection Phase (ASP). These phases are visualized in Figure 1. Each test case is going through these phases individually, except STP, which uses the results of the previous phase to evaluate the hypothesis.

The following list provides a general description of each phase, and clarifies its purpose:

- Preprocessing Phase. The purpose of this phase is to make the data sets less complex and faster to classify.
  - Common Preprocessing Phase (CPP). The purpose of this phase is to clean the data set and reduce its size. We do this by removing grammatical elements and combining similar textual elements.
  - Varying Preprocessing Phase (VPP). This phase applies the preprocessing methods that we want to test.
It consists of multiple preprocessing steps, which are continuously changed based on the VPP configuration of the test case.

- **Attribute Selection Phase (ASP).** During this phase, we evaluate the preprocessed data set and remove attributes from them in order to reduce classification time.
- **Classification Phase (CP).** During this phase, we use the preprocessed data set to train and test a classifier.
- **Statistical Test Phase (STP).** During this phase, we use a statistical test on the results gained from CP to evaluate our hypothesis.

**B. Our Application of the Framework**

We use a set of Danish tweets and a set of English tweets as the input set for our framework.

Following is a description of the specific methods used for our implementation of each of the phases described in our framework:

1) **Common Preprocessing Phase (CPP):** Our input for this phase consists of several preprocessing methods which are described below (in execution order):

- **Replace user**
  Replaces a mention of a user, e.g. ‘@johndoe’ with ‘<user>’ in order to unify all references to users.
- **Replace link**
  Replaces a link, e.g. ‘pic.twitter.com’ with ‘<link>’ in order to unify all references to links, since we do not want to distinguish between links.
- **Remove repeated characters**
  Repeated characters in a word are reduced to a maximum of three repetitions. For example, the word ‘happpppyyyy’ becomes ‘happyyyy’. This is done because a maximum of two adjacent character repetitions can occur naturally, and we assume there is little intensity difference based on the exact amount of repetitions (e.g. ‘saaad’ and ‘saaaaad’ have roughly the same intensity). However, we expect a substantial intensity difference between using repeated characters and not which is why up to three repetitions are kept (e.g. ‘sad’ and ‘saaad’ have different intensities).
- **Hashtag deletion**
  Hashtags are replaced with the word in the hashtag, e.g. ‘#sad’ becomes ‘sad’. Hashtags are often included as words in the text or to summarize the tweet, which is the reason they are kept.
- **Replace emoticons**
  Each emoticon is replaced with an equivalent emoji, thereby reducing the number of attributes. For example, ‘:D’ and ‘:-D’ both become ‘😊’.
- **Lowercasing**
  All tweets are converted to lowercase.
- **Symbol removal**
  All symbols are removed from the tweets. Commas and semicolons are replaced with ‘<soft>’, and additionally ‘<soft>’ is also added after every string of emojis. Dots, colons, exclamation marks, and question marks are replaced with ‘<hard>’.
  ‘<soft>’ and ‘<hard>’ are later used in ‘n-gram stop-split’ step, described in VPP.

2) **Varying Preprocessing Phase (VPP):** Our input for this phase consists of the following preprocessing methods (described in execution order):

- **Part-Of-Speech (POS) tagger**
  A POS tagger finds the corresponding word class for each word in the data sets. This is done to focus on typical emotional word classes, i.e. nouns, adjectives, adverbs, and verbs, by removing words from all other classes[8].
- **Stemming**
  Stemming is a process, where each word is converted to its root (e.g. ‘walking’ becomes ‘walk’ and ‘smiling’
In the calculating expected entropy we reduce the domain of all information gain of each attribute, and remove all attributes to whether the word is in the tweet or not.

\[ \text{Gain}(X) = h(C) - h(C|X) \]

\[ h(C) = \sum_{i=1}^{n} -C_i \cdot \log_2(C_i) \]

\[ h(C|X) = \sum_{i=1}^{m} \frac{|E_i|}{|E|} \cdot h_i(C) \]

In these equations \( C \) is the set of classes \( \{C_1, C_2, \ldots, C_n\} \), where \( C_i \) refers to a specific class, \( X \) is an attribute with the domain \( X = \{v_1, v_2, \ldots, v_m\} \), where \( v_i \) refers to a specific value in the domain, \( E_i \) is the set of instances with \( X = v_i \), and \( h_i(C) \) is the entropy of classes in \( E_i \).

The domain of our attributes describes how many times the \( n \)-gram is used in a tweet. However, for the purposes of calculating expected entropy we reduce the domain of all attributes to whether the word is in the tweet or not.

4) Classification Phase (CP): We run all our classifiers using Weka\(^1\). In order to minimize bias and randomness, we use Weka’s standard parameters, with a 5 fold cross-validation. Which classification algorithm is used depends on the classification configuration from the following options:

- Support-Vector Machine (SVM) - A nonprobabilistic binary classification algorithm. It constructs a hyperplane to separate two classes based on the data points closest to the gap between the classes. We use the SVM optimizer Sequential Minimal Optimization (SMO) for this[10][11].
- Random Forest (RF) - It is also known as random decision forest. RF generates random decision trees which can be used for classification, regression and other purposes[12].
- Naive Bayes (NB) - A simple probabilistic classification algorithm based on applying Bayes’ theorem with strong independence assumptions between the features[13].

5) Statistical Test Phase (STP): For our STP, we use a two-sided Wilcoxon signed-rank test[14] on the accuracy difference in pairs of test cases across languages in order to test the following hypothesis:

**Hypothesis:** The change in classification accuracy for Emotional Classification caused by changing a single preprocessor or classifier is independent of the target language within a significance level of \( p = 0.05 \).

We cannot use the raw accuracy difference between the languages, since that will only show the difference in difficulty of doing EC on the two languages. Instead we calculate the difference between pairs of classification configurations using our classification results. The difference between the classification configuration pair \((A, B)\) is calculated as: \( A - B \). We create a pair of test case configurations \((A, B)_{\text{Danish}}, (A, B)_{\text{English}}\) consisting of two pairs of classification configurations.

The test cases representing this test case configuration pair are used as a pair of data points for the Wilcoxon test to make our cross-language comparison. We do this for each pair of classification configurations \((A, B)\) which only have one difference between them (one varying preprocessor or a different classifier), making up a total of 180 pairs of data points for the Wilcoxon test. These pairs of data points can be seen in Table IV.

We are not using pairs of classification configurations with more than one difference between them since they are already represented through multiple pairs of classification configurations with only one difference: \((A - C) = (A - B) + (B - C)\).

V. Experiment

During this Section, we specify some details of our experiment, specifically our data extraction process and VPP configurations. We conduct this experiment in order to determine whether the language being classified has impact on the accuracy of EC for a given classification configuration or not.

---

\(^1\)https://www.cs.waikato.ac.nz/~ml/weka/
A. VPP Configurations

All possible VPP configurations for our experiment are shown in Table I. We use these configurations both for the Danish and the English data set, and for each classifier. This table uses the following abbreviations for describing the types of VPP methods included in each VPP configuration:

- Stemming = ST
- POS tagger = POS
- n-gram = NG
- n-gram stop-split = NGSS

B. Data Extraction

For each base emotion, we manually choose hashtags based on synonyms and similar words from these websites\(^2\)\(^3\)\(^4\). Then we manually filter the hashtags, based on whether the tweets using the hashtag show the correct emotion. Examples of these hashtags are shown in Table II. We then download the tweets, which include the remaining hashtags, using the python library `Twin`\(^5\).

It is important that the data set for each language are as similar as possible. This is to ensure that any difference we detect in the performance of methods is due to linguistic differences rather than other differences in the data sets. In particular, we want the data sets to have equal size and distribution between classes. The English data set is created based on the size of the Danish data set since there are fewer Danish tweets compared to English tweets. For each English hashtag, we collected a number of tweets equal to \(\frac{1}{2}\) of the number of Danish tweets for the class which the hashtag belongs to. Then from each class of English tweets a number of random unique tweets, equal to the size of the same class of Danish tweets, are selected. This makes the data sets equal in number of tweets for each class, as well as in the total number of tweets.

VI. Evaluation

In this Section, we show and discuss the results from our experiment’s CP and STP through trends and phenomena that occur.

A. Classification Evaluation

For each test case configuration, we calculate accuracy, precision, recall, and F-measure using Weka. Accuracy is a general measure of the quality of the classification. Precision and recall are both measures of relevance, where precision describes how many retrieved items are relevant, and recall describes how many relevant items are retrieved. The values listed are the average precision and recall of the classes. F-measure is the harmonic mean of precision and recall. The values listed are the average F-measure of the classes. Weka calculates these statistics using the following formulas:

\[
\text{Accuracy} = \frac{|\text{correct results}|}{|\text{correct results} \cup \text{incorrect results}|} \quad (4)
\]

\[
\text{Precision} = \frac{1}{n} \sum_{i=1}^{n} \frac{|TP(C_i)|}{|TP(C_i)| + |FP(C_i)|} \quad (5)
\]

\[
\text{Recall} = \frac{1}{n} \sum_{i=1}^{n} \frac{|TP(C_i)|}{|TP(C_i)| + |FN(C_i)|} \quad (6)
\]

\[
F\text{-measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)
\]

In the above equations, \(C\) is the set of classes \(C = \{C_1, C_2, \ldots, C_n\}\), where \(C_i\) refers to a specific class of emotions, and \(n\) is the number of classes (eight emotions in our case). correct results is the set of all results which are classified as the correct class, while incorrect results is the set of all results classified as the wrong class. \(TP(C_i), TN(C_i), FP(C_i),\) and \(FN(C_i)\) describe the set of: true positive-, true negative-, false positive-, and false negative results respectively, for the class \(C_i\).

We present the results of the CP in Table III. A row in Table III describes which VPP configuration is used, while the columns describe whether accuracy, F-measure, precision, or recall is shown, and which language and classification algorithm is used.

When we observe the results, the following trends appear:

- The average accuracy of the English data set is lower than the average accuracy of the Danish data set. This might be due to the higher diversity in English tweets, created by the difference in numbers of hashtags, and that the English tweets are written by different cultures.

---

2https://ordnet.dk/
3https://www.thesaurus.com/
4https://sproget.dk/dk/
5https://github.com/twinproject/twin

<table>
<thead>
<tr>
<th>Configuration #</th>
<th>Configuration Setup</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>(NG = {1})</td>
</tr>
<tr>
<td>C2</td>
<td>(NG = {1, 2})</td>
</tr>
<tr>
<td>C3</td>
<td>(NG = {1, 2, 3})</td>
</tr>
<tr>
<td>C4</td>
<td>(NG = {1}, ST)</td>
</tr>
<tr>
<td>C5</td>
<td>(NG = {1, 2}, ST)</td>
</tr>
<tr>
<td>C6</td>
<td>(NG = {1, 2, 3}, ST)</td>
</tr>
<tr>
<td>C7</td>
<td>(NG = {1}, POS)</td>
</tr>
<tr>
<td>C8</td>
<td>(NG = {1, 2}, POS)</td>
</tr>
<tr>
<td>C9</td>
<td>(NG = {1, 2, 3}, POS)</td>
</tr>
<tr>
<td>C10</td>
<td>(NG = {1, 2}, NGSS)</td>
</tr>
<tr>
<td>C11</td>
<td>(NG = {1, 2, 3}, NGSS)</td>
</tr>
<tr>
<td>C12</td>
<td>(NG = {1}, ST, POS)</td>
</tr>
<tr>
<td>C13</td>
<td>(NG = {1, 2}, ST, POS)</td>
</tr>
<tr>
<td>C14</td>
<td>(NG = {1, 2, 3}, ST, POS)</td>
</tr>
<tr>
<td>C15</td>
<td>(NG = {1, 2}, POS, NGSS)</td>
</tr>
<tr>
<td>C16</td>
<td>(NG = {1, 2, 3}, POS, NGSS)</td>
</tr>
<tr>
<td>C17</td>
<td>(NG = {1, 2}, ST, NGSS)</td>
</tr>
<tr>
<td>C18</td>
<td>(NG = {1, 2, 3}, ST, NGSS)</td>
</tr>
<tr>
<td>C19</td>
<td>(NG = {1, 2}, ST, POS, NGSS)</td>
</tr>
<tr>
<td>C20</td>
<td>(NG = {1, 2, 3}, ST, POS, NGSS)</td>
</tr>
</tbody>
</table>
TABLE II
EXAMPLES AND NUMBER OF HASHTAGS AND TWEETS.

<table>
<thead>
<tr>
<th>Emotion</th>
<th># of Hashtags</th>
<th>Danish Hashtag Examples</th>
<th># of Tweets</th>
<th>Danish Tweet Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joy</td>
<td>25</td>
<td>#glad #glade #fryd</td>
<td>16541</td>
<td>Hold nu op hvor jeg elsker faneblade i Finder i OSX. Det er SÅ genialt! #glade</td>
</tr>
<tr>
<td>Trust</td>
<td>16</td>
<td>#bulls #bullsford #bitro</td>
<td>4125</td>
<td>Nu synes jeg godt snart det må falde lidt til ro i Japan tak! #bekymret</td>
</tr>
<tr>
<td>Fear</td>
<td>35</td>
<td>#fryg #angst #bange</td>
<td>4941</td>
<td>Så har man set det med. Unge tabere der leger med lasere... #chokeret</td>
</tr>
<tr>
<td>Surprise</td>
<td>26</td>
<td>#overrasket #florandet #forbavset</td>
<td>2224</td>
<td>Ø, hvor kan man nogen gange blive lidt trist til mode, over de minstede ting #test</td>
</tr>
<tr>
<td>Sadness</td>
<td>33</td>
<td>#ked #kedafdet #deprimeret</td>
<td>20537</td>
<td>Er et skridt tættere på at være voksen efter jeg har renset afløb i mit badeværelse! #ulækkert</td>
</tr>
<tr>
<td>Disgust</td>
<td>39</td>
<td>#beskidt #snavset #gyselig</td>
<td>3889</td>
<td>Jeg håber at der er en der saver Suarez fuldstændigt midt over. #bitter</td>
</tr>
<tr>
<td>Anger</td>
<td>34</td>
<td>#vred #arrig #hidsig</td>
<td>4056</td>
<td>Jeg fucking håber Lady Gaga kommer til Danmark! #tab</td>
</tr>
<tr>
<td>Anticipation</td>
<td>20</td>
<td>#spændende #nygerrig #fristende</td>
<td>8859</td>
<td>Save the date! Nov 9th to 16th! #expectation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Emotion</th>
<th># of Hashtags</th>
<th>English Hashtag Examples</th>
<th># of Tweets</th>
<th>English Tweet Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joy</td>
<td>39</td>
<td>#joy #happy #happiness</td>
<td>16541</td>
<td>Final week of semester! #contented</td>
</tr>
<tr>
<td>Trust</td>
<td>15</td>
<td>#trust #trustful #admiration</td>
<td>4125</td>
<td>Don’t #depend on others when you can #doityourself!</td>
</tr>
<tr>
<td>Fear</td>
<td>49</td>
<td>#fear #error #right</td>
<td>4941</td>
<td>one of the #worst features about #worrying is that it destroys our ability to #centerate.</td>
</tr>
<tr>
<td>Surprise</td>
<td>25</td>
<td>#surprise #surprising #amazement</td>
<td>2224</td>
<td>@netflix love death &amp; robots is amazing, loving it #astonishing</td>
</tr>
<tr>
<td>Sadness</td>
<td>47</td>
<td>#grief #sadness #sorrow</td>
<td>20537</td>
<td>I am not sure I care anymore #painful</td>
</tr>
<tr>
<td>Disgust</td>
<td>69</td>
<td>#ew #unclean #jealous</td>
<td>3889</td>
<td>I went back to high school for two hours and that’s time I can never get back #resent #regret</td>
</tr>
<tr>
<td>Anger</td>
<td>43</td>
<td>#angry #anger #mad</td>
<td>4056</td>
<td>I hate Iowa #displeased</td>
</tr>
<tr>
<td>Anticipation</td>
<td>28</td>
<td>#anticipation #watchful #expecting</td>
<td>8859</td>
<td>Save the date! Nov 9th to 16th! #expectation</td>
</tr>
</tbody>
</table>

while the Danish tweets primarily are written by Danish people.

• SVM has the highest accuracy, F-measure, precision, and recall, out of all classifiers and across both languages.

• The n-gram stop-split preprocessor does not make a large difference in the results. There are only a few cases with a noticeable difference, e.g. between C16 and C9, which is NG = \{1, 2, 3\} POS, with and without NGSS respectively. This might be because most of the n-grams in this preprocessor removes would otherwise have been removed during the ASP.

• The differences in classification effectiveness between NG = \{1\} and NG = \{1, 2, 3\} is the opposite of what we expected. The effectiveness of NG = \{1\} is often higher than the other n-gram variations for both Danish and English. This suggests that the context gained from adding orders of words is less significant than the noise created by adding more n-gram attributes.

B. Statistical Test Evaluation

We compare the classification accuracies, from Table III by applying them on a Wilcoxon test. The basis of this analysis is described in Section IV-B5.

Figure 2 shows the pairs of test case configurations where Table IV shows the setup of each test case configuration i.e. the variables on the x-axis of Figure 2.

In Table IV, test case configuration differences written on the form ‘VPP configuration-classifier-classifier’ describe two test case configurations with the same VPP configuration but different classifiers. However, test case configuration differences on the form ‘VPP configuration-VPP configuration-classifier’ describe two test case configurations with one difference in their VPP configuration but using the same classifier. The corresponding VPP configurations are shown in Table I.

In Figure 2, each point represents the difference between two test cases’ accuracy (A_{accuracy}, B_{accuracy}), where A and B has only one difference between their classification configurations. If a point is positive, then test case A has a higher accuracy than B; if a point is negative, then test case A has a lower accuracy than B; and if a point is 0, then there is no difference between their accuracies.

Each line in Figure 2 represents the accuracy difference between a pair of test case configurations. The red and orange lines represent the English data set, while the blue and cyan represents the Danish data set. The special cases where one point is above 0 and the other is below, represent test case configuration pairs where there is a positive accuracy change for one language and a negative change for the other. Orange and cyan represent these special cases. These cases support the rejection of our hypothesis.

Running the Wilcoxon test on our test case configuration pairs results in a p-value of 0.1285. Our hypothesis is therefore not rejected within a significance level of 0.05. Thus, which classification configuration that performs best might be independent of the languages being classified.

The box plot in Figure 3 shows the variance of the accuracy difference in the data used for the Wilcoxon test. We can see that the English data set has a higher variance, meaning it is more sensitive towards configuration changes. Despite this, both data sets have a median close to 0, which could explain why we cannot reject our hypothesis.

By studying Figure 2, we learn that the biggest differences in accuracy comes from the change of classifier to/from NB. Furthermore, POS tagging on the English data set makes almost
as large a negative change in accuracy difference as changing classifier to NB. The Danish data set however improves slightly when POS tagging is applied. This effect can be seen in the difference between C1 to C7, C2 to C8, and C3 to C9 for all classifiers.

### VII. DISCUSSION

In this Section, we discuss the consequences of the observations in Section VI-B. First we look at the results of the Wilcoxon test, followed by the effects of classifiers, and language specific tools.

As described in Section VI-B, our Wilcoxon test did not yield any significant results. This suggests that classification configurations react similarly to the Danish and the English data set. However, further research is needed to establish the statement “EC research based on one language is applicable to other languages”.

However, there is a significant difference when NB is applied as a classifier. Using NB, the accuracies of the English data set are between 50% – 58% while the Danish data set’s accuracies are between 72% – 80%. This suggests that there is a relevant difference in EC between the two languages.
Fig. 2. Data points used in the Wilcoxon test. Configurations can be seen in Table IV. Each data point represents the percentage difference in accuracy between a pair of classification configurations. Data points marked with a blue circle represents the Danish data set and points marked with a red square represents the English data set. The data points with a orange diamonds and cyan triangles represent special cases for Danish and English respectively. These special cases describe where the configuration change had a positive impact on the one language but not with the other.

Another interesting observation we found in Section VI-B is that POS tagging has opposite effects on the two languages. Adding POS tagging made a difference in accuracy between $-0.58\%$ and $7.02\%$ on the Danish data set and between $-1.82\%$ and $-25.78\%$ on the English data set. The variance is not only higher for the English data set, as shown in Figure 3, the difference is also mostly positive for Danish and always negative for English. This means that the Danish data set benefits from POS tagging while the English data set suffers greatly from it. This suggests that while a lot of the elements of EC are not language dependent, the use of tools designed for a specific language might be language dependent. Therefore, more language specific research in these tools would be beneficial.

A. Possible Error Sources

By analyzing our experiment, we find some possible error sources which may have impact on our results.

- There exists non-Danish tweets in the Danish data set since Twitter’s language filter is not perfect.
- English tweets are posted more often than Danish tweets, and we download the tweets in chronological descending order of posting time. In order to have the same amount of tweets in the data sets, the Danish data set ends up with a much higher time variance between posts than the English data set. Therefore, the Danish data set probably has a higher variance in how the language is used.
- The hashtags used for gathering tweets have been chosen manually and therefore do not cover all emotional words related to the base emotions.
- There may be differences in how the chosen hashtags are related to the base emotion they are labeled with. There are also 87 more English hashtags than Danish hashtags. This might cause the English data set to be more diverse and therefore possibly harder to classify.
- There are some differences between the Danish and
English POS tagging and stemming preprocessing methods used in VPP.

- The Danish POS tagger labels nonwords as nouns and the English POS tagger labels nonwords as proper nouns. In the VPP the POS tagging preprocessor keeps nouns but not proper nouns as part of the attributes.

### VIII. Conclusion

We have conducted this study in order to test whether the new research field: cross-language EC has the potential for reducing the amount of research needed for non-English languages within the field of EC. In Section IV-A, we constructed a framework for testing the classification accuracy of a number of test cases. In Section IV-B, the framework was used to setup our experiment, for the purpose of evaluating our null hypothesis: The change in classification accuracy for Emotional Classification caused by changing a single preprocessor or classifier is independent of the target language within a significance level of $p = 0.05$. We made this hypothesis in order to answer the more general question: Do target languages have impact on the effectiveness of EC methods? Our two-sided Wilcoxon signed-rank test gave a $p$-value of 0.12852, and therefore did not reject the hypothesis using data sets constructed from Danish and English tweets. It should be noted that our results are based only on two germanic languages with the common domain Twitter, and thus only covers a small part of the research within cross-language EC. During our experiment, SVM has consistently yielded the best results in contrast to the experiment made by [6], where SVM did not yield consistent results on nonbinary classification. In Section VI, we observed a few interesting characteristics of our results, e.g. POS tagging works well for the Danish data set but not for the English data set. These findings suggest that further research is needed for cross-language EC. We believe our study

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1 C1-SVM-NB</td>
<td>31 C11-SVM-NB</td>
<td>61 C1-C2-SVM</td>
<td>91 C4-C5-SVM</td>
<td>121 C8-C13-SVM</td>
<td>151 C13-C14-SVM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 C1-SVM-RF</td>
<td>32 C11-SVM-RF</td>
<td>62 C1-C2-RF</td>
<td>92 C4-C5-RF</td>
<td>122 C8-C13-RF</td>
<td>152 C13-C14-RF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 C2-SVM-NB</td>
<td>33 C11-NB-RF</td>
<td>63 C1-C2-RF</td>
<td>93 C4-C5-RF</td>
<td>123 C8-C13-RF</td>
<td>153 C13-C14-RF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 C2-SVM-RF</td>
<td>34 C12-SVM-NB</td>
<td>64 C1-C4-SVM</td>
<td>94 C4-C12-SVM</td>
<td>124 C8-C15-SVM</td>
<td>154 C13-C19-SVM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 C2-SVM-RF</td>
<td>35 C12-SVM-RF</td>
<td>65 C1-C4-NB</td>
<td>95 C4-C12-RF</td>
<td>125 C8-C15-RF</td>
<td>155 C13-C19-NB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 C2-NB-RF</td>
<td>36 C12-NB-RF</td>
<td>66 C1-C4-RF</td>
<td>96 C4-C12-RF</td>
<td>126 C8-C15-RF</td>
<td>156 C13-C19-RF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 C3-SVM-RF</td>
<td>37 C13-SVM-NB</td>
<td>67 C1-C7-SVM</td>
<td>97 C5-C6-SVM</td>
<td>127 C9-C14-SVM</td>
<td>157 C14-C20-SVM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 C3-SVM-RF</td>
<td>38 C13-SVM-RF</td>
<td>68 C1-C7-NB</td>
<td>98 C5-C6-NB</td>
<td>128 C9-C14-NB</td>
<td>158 C14-C20-NB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 C3-NB-RF</td>
<td>39 C13-NB-RF</td>
<td>69 C1-C7-RF</td>
<td>99 C5-C6-RF</td>
<td>129 C9-C14-RF</td>
<td>159 C14-C20-RF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 C3-SVM-NB</td>
<td>40 C14-SVM-NB</td>
<td>70 C2-C3-SVM</td>
<td>100 C5-C13-SVM</td>
<td>130 C9-C16-SVM</td>
<td>160 C15-C16-SVM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 C4-SVM-RF</td>
<td>41 C14-SVM-RF</td>
<td>71 C2-C3-NB</td>
<td>101 C5-C13-NB</td>
<td>131 C9-C16-NB</td>
<td>161 C15-C16-NB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12 C4-NB-RF</td>
<td>42 C14-NB-RF</td>
<td>72 C2-C3-RF</td>
<td>102 C5-C13-RF</td>
<td>132 C9-C16-RF</td>
<td>162 C15-C16-RF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13 C5-SVM-NB</td>
<td>43 C15-SVM-NB</td>
<td>73 C2-C5-SVM</td>
<td>103 C5-C17-SVM</td>
<td>133 C10-C11-SVM</td>
<td>163 C15-C19-SVM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14 C5-SVM-RF</td>
<td>44 C15-SVM-RF</td>
<td>74 C2-C5-NB</td>
<td>104 C5-C17-NB</td>
<td>134 C10-C11-NB</td>
<td>164 C15-C19-NB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 C5-NB-RF</td>
<td>45 C15-NB-RF</td>
<td>75 C2-C5-RF</td>
<td>105 C5-C17-RF</td>
<td>135 C10-C11-RF</td>
<td>165 C15-C19-RF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16 C6-SVM-NB</td>
<td>46 C16-SVM-NB</td>
<td>76 C2-C8-SVM</td>
<td>106 C6-C14-SVM</td>
<td>136 C10-C15-SVM</td>
<td>166 C16-C20-SVM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17 C6-SVM-NB</td>
<td>47 C16-SVM-RF</td>
<td>77 C2-C8-NB</td>
<td>107 C6-C14-NB</td>
<td>137 C10-C15-NB</td>
<td>167 C16-C20-NB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 C6-NB-RF</td>
<td>48 C16-NB-RF</td>
<td>78 C2-C8-RF</td>
<td>108 C6-C17-RF</td>
<td>138 C10-C15-RF</td>
<td>168 C16-C20-RF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19 C7-SVM-NB</td>
<td>49 C17-SVM-NB</td>
<td>79 C2-C10-SVM</td>
<td>109 C6-C18-SVM</td>
<td>139 C10-C17-SVM</td>
<td>169 C17-C18-SVM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 C7-SVM-RF</td>
<td>50 C17-SVM-RF</td>
<td>80 C2-C10-RF</td>
<td>110 C6-C18-NB</td>
<td>140 C10-C17-NB</td>
<td>170 C17-C18-NB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21 C7-NB-RF</td>
<td>51 C17-NB-RF</td>
<td>81 C2-C10-RF</td>
<td>111 C6-C18-RF</td>
<td>141 C10-C17-RF</td>
<td>171 C17-C18-RF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22 C8-SVM-NB</td>
<td>52 C18-SVM-NB</td>
<td>82 C3-C6-SVM</td>
<td>112 C7-C8-SVM</td>
<td>142 C11-C16-SVM</td>
<td>172 C17-C19-SVM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23 C8-SVM-RF</td>
<td>53 C18-SVM-RF</td>
<td>83 C3-C6-NB</td>
<td>113 C7-C8-NB</td>
<td>143 C11-C16-NB</td>
<td>173 C17-C19-NB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24 C8-NB-RF</td>
<td>54 C18-NB-RF</td>
<td>84 C3-C6-RF</td>
<td>114 C7-C8-RF</td>
<td>144 C11-C16-RF</td>
<td>174 C17-C19-RF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26 C9-SVM-RF</td>
<td>56 C19-SVM-RF</td>
<td>86 C3-C9-NB</td>
<td>116 C7-C12-NB</td>
<td>146 C11-C18-NB</td>
<td>176 C18-C20-NB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27 C9-NB-RF</td>
<td>57 C19-NB-RF</td>
<td>87 C3-C9-RF</td>
<td>117 C7-C12-RF</td>
<td>147 C11-C18-RF</td>
<td>177 C18-C20-RF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29 C10-SVM-RF</td>
<td>59 C20-SVM-RF</td>
<td>89 C3-C11-NB</td>
<td>119 C8-C9-NB</td>
<td>149 C12-C13-NB</td>
<td>179 C19-C20-NB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30 C10-NB-RF</td>
<td>60 C20-NB-RF</td>
<td>90 C3-C11-RF</td>
<td>120 C8-C9-RF</td>
<td>150 C12-C13-RF</td>
<td>180 C19-C20-RF</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
is significant as it introduces a new topic within EC with the potential to help other EC research.

A. Future Work

The experiment we have conducted is only a small part of cross-language classification research since it only tested on the Danish and English language, a few preprocessing methods, and three classification algorithms. Therefore it is necessary to make similar experiments, e.g. on languages other than Danish and English in order to validate our hypothesis. Researching the cross-language effectiveness of other preprocessors and classifiers is also a possible continuation of our work. It will also be worth testing the differences between languages with different alphabets and/or structure, especially Latin-based and non-Latin-based languages. The framework described in Section IV can serve as a guide for comparing EC methods between languages. Whether languages have impact on the effectiveness of preprocessing and classification methods is still an open problem, that can be tested using other languages, preprocessing methods, classification algorithms, and/or data sets. One possible data set to use would be the SemEval-2019 data set6, which is used for a semantic evaluation workshop.

REFERENCES


6http://alt.qcri.org/semeval2019/
