Aalborg Universitet

AALBORG UNIVERSITY

Languages' impact on emotional classification methods

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

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 Lagerløfs 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 Statical 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 Naive 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 Naive Bayes (NB) as it is a common classification algorithm. We also use 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.
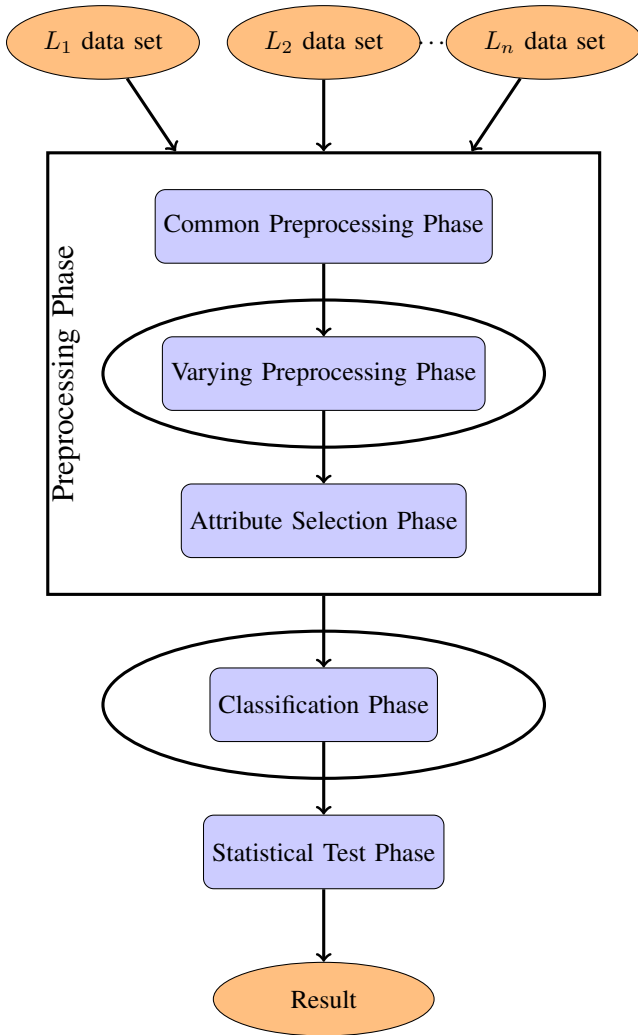
Fig. 1. The process of our framework. L_1 to L_n represents a minimum of two data sets in different languages to be tested. The black box describes the preprocessing phase which involves the following subphases: CPP, VPP, and ASP. The black ellipses describe the varying parts of our experiment which are changed for each test case.

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.
- Statical 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 'happppppyyyy' becomes 'happpyyy'. 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'

becomes 'smile'). While some intensity may be lost, the number of attributes are greatly reduced.

- **n-gram stop-split**
  In this step the *<soft>* and *<hard>* stops are used to split tweets into multiple sets of words, which are split further by n-gram before being classified. This means that conjunctions and interposed sentences are taken into account when classifying longer sentences. We use this preprocessor in order to account for the difference in the use of commas between the Danish and the English language. The varying part here is whether *<soft>* is used to split tweets or not while *<hard>* is always used to find splits.

- **n-gram**
  n-gram splits the sets of words acquired in the n-gram stop-split preprocessor into smaller sets of words. We test $NG = \{1\}$, $NG = \{1, 2\}$, and $NG = \{1, 2, 3\}$ n-gram combination since combinations of multiple n-grams received better results than single n-grams in [4].

3) Attribute Selection Phase (ASP): Our input for this phase consists of two different methods for removing attributes. Firstly, attributes that only appear in the data set once are removed because they cannot be in the test set and training set at the same time. Besides this we also evaluate the information gain of each attribute, and remove all attributes with an information gain less than 0.00025. This reduced the number of attributes substantially, e.g. for our test case with the most attributes, $NG = \{1, 2, 3\}\ English$, we started with $1,719,816$ attributes, and after running the ASP it had $15,210$ attributes left.

Information gain describes how much information an attribute gives us about the classes. It is calculated using Equation 1, which uses Equation 2 and Equation 3 describing entropy and expected entropy respectively[9].

$$Gain(X) = h(C) - h(C|X) \quad (1)$$

$$h(C) = \sum_{i=1}^{n} -C_i \cdot log_2(C_i) \quad (2)$$

$$h(C|X) = \sum_{i=1}^{m} \frac{|E_i|}{|E|} \cdot h_i(C) \quad (3)$$

In these equations C is the set of classes $C = \{C_1, C_2, \ldots, C_n\}$, 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) Statical 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 $((A, B)_{Danish}, (A, B)_{English})$ 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/

TABLE I
VPP CONFIGURATIONS USED IN OUR EXPERIMENT. CONFIGURATIONS WITH $NG = \{1\}$ AND NGSS ARE REMOVED AS N-GRAM STOP-SPLIT HAS NO IMPACT ON 1-GRAMS.

Configuration #	Configuration Setup
C1	$NG = \{1\}$
C2	$NG = \{1, 2\}$
C3	$NG = \{1, 2, 3\}$
C4	$NG = \{1\}$, ST
C5	$NG = \{1, 2\}$, ST
C6	$NG = \{1, 2, 3\}$, ST
C7	$NG = \{1\}$, POS
C8	$NG = \{1, 2\}$, POS
C9	$NG = \{1, 2, 3\}$, POS
C10	$NG = \{1, 2\}$, NGSS
C11	$NG = \{1, 2, 3\}$, NGSS
C12	$NG = \{1\}$, ST, POS
C13	$NG = \{1, 2\}$, ST, POS
C14	$NG = \{1, 2, 3\}$, ST, POS
C15	$NG = \{1, 2\}$, POS, NGSS
C16	$NG = \{1, 2, 3\}$, POS, NGSS
C17	$NG = \{1, 2\}$, ST, NGSS
C18	$NG = \{1, 2, 3\}$, ST, NGSS
C19	$NG = \{1, 2\}$, ST, POS, NGSS
C20	$NG = \{1, 2, 3\}$, ST, POS, NGSS

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 'Twint'[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

[2]https://ordnet.dk/
[3]https://www.thesaurus.com/
[4]https://sproget.dk/
[5]https://github.com/twintproject/twint

hashtag, we collected a number of tweets equal to $\frac{1}{10}$ 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:

$$Accuracy = \frac{|correct\ results|}{|\ correct\ results \cup incorrect\ results\ |} \quad (4)$$

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

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

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

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 the higher diversity in English tweets, created by the difference in numbers of hashtags, and that the English tweets are written by many different cultures,

TABLE II
EXAMPLES AND NUMBER OF HASHTAGS AND TWEETS.

Emotion	# of Hashtags	Danish Hashtag Examples	# of Tweets	Danish Tweet Example
Joy	25	#glad #glæde #fryd	16541	Hold nu op hvor jeg elsker faneblade i Finder i OSX. Det er SÅ genialt! #glæde
Trust	16	#tillid #tillidsfuld #tiltro	4125	Når en fyr viser han er til at stole på #tillid
Fear	35	#frygt #angst #bange	4941	Nu synes jeg godt snart det må falde lidt til ro i Japan tak! #Bekymret
Surprise	26	#overrasket #forundret #forbavset	2224	Så har man set det med.. Unge tabere der leger med lasere... #chokeret
Sadness	33	#ked #kedafdet #deprimeret	20537	Øv, hvor kan man nogen gange blive lidt trist til mode, over de mindste ting #trist
Disgust	39	#beskidt #snavset #gyselig	3889	Er et skridt tættere på at være voksen efter jeg har renset afløb i mit badeværelse! #ulækkert
Anger	34	#vred #arrig #hidsig	4056	Jeg håber at der er en der saver Suarez fuldstændig midt over. #bitter
Anticipation	20	#spændende #nysgerrig #fristende	8859	Jeg fucking håber Lady Gaga kommer til Danmark! #håb

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

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 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 0.12852. 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

TABLE III
TEST CASE RESULTS: BOLD VALUES ARE THE HIGHEST VALUES WITHIN THE CLASSIFIER AND LANGUAGE COMBINATION WHILE UNDERLINED VALUES ARE THE HIGHEST VALUES WITHIN THE LANGUAGE.

Config.	Accuracy						F-measure					
	Danish			English			Danish			English		
	SVM	NB	RF	SVM	NB	RF	SVM	NB	RF	SVM	NB	RF
C1	94.66	76.45	91.63	**94.24**	55.58	87.14	98.33	79.09	96.94	**95.54**	58.68	93.14
C2	94.45	75.40	89.80	93.97	54.39	83.20	98.32	78.33	96.36	95.23	57.60	90.99
C3	94.40	75.55	88.98	93.88	54.35	81.69	98.30	78.55	96.01	95.19	57.56	89.92
C4	93.97	73.84	90.04	92.81	**57.36**	83.61	97.89	77.85	94.38	94.38	**60.97**	90.30
C5	93.68	72.93	87.85	92.48	56.09	80.57	97.94	77.33	95.27	94.11	59.84	88.74
C6	93.66	73.04	86.95	92.46	55.94	79.30	97.95	77.33	94.86	94.12	59.67	87.92
C7	**94.71**	78.59	**92.95**	68.46	51.10	67.56	**98.35**	81.16	**97.52**	69.26	52.00	68.63
C8	94.54	77.98	92.22	69.01	50.95	67.85	98.33	81.02	97.33	69.86	52.51	69.17
C9	94.53	77.93	91.77	68.98	51.00	67.73	98.31	81.07	97.25	69.66	52.54	69.28
C10	94.45	75.29	89.56	94.01	54.60	83.52	98.34	78.17	96.23	95.26	57.83	91.00
C11	94.40	75.23	88.90	93.93	54.78	**91.73**	98.31	78.21	95.90	95.26	57.98	**97.12**
C12	93.40	**79.71**	91.65	67.55	54.56	67.42	97.76	**83.31**	97.14	68.38	54.59	68.56
C13	93.20	79.01	90.59	68.48	54.29	67.68	97.89	82.99	96.94	69.23	54.91	69.14
C14	93.14	78.92	90.23	68.44	54.11	67.38	97.89	82.99	96.83	69.14	54.83	68.85
C15	94.55	77.98	92.18	69.03	51.04	67.88	98.32	81.02	97.27	69.68	52.34	69.24
C16	94.52	77.91	81.88	69.01	51.02	67.81	98.32	81.08	90.03	69.62	52.43	69.31
C17	93.71	72.83	87.53	92.55	56.46	80.78	97.93	76.93	95.22	94.19	60.23	89.02
C18	93.66	72.83	86.86	92.51	56.63	79.63	97.94	76.98	94.82	94.12	60.42	88.21
C19	93.22	79.01	90.51	68.33	54.38	67.57	97.88	82.98	96.89	68.99	54.84	68.80
C20	93.17	78.92	90.22	68.45	54.14	67.54	97.88	82.98	96.72	69.08	54.72	68.94
Avg.	94.00	76.46	89.62	80.92	54.14	75.38	98.11	79.97	96.08	82.01	56.32	79.81

Config.	Precision						Recall					
	Danish			English			Danish			English		
	SVM	NB	RF	SVM	NB	RF	SVM	NB	RF	SVM	NB	RF
C1	98.22	74.80	95.98	**95.36**	50.65	90.21	98.44	83.94	97.91	95.72	69.77	96.25
C2	98.35	73.98	95.24	94.68	49.20	87.61	98.30	83.24	97.51	95.78	69.47	94.65
C3	98.33	74.39	94.69	94.62	49.02	86.31	98.27	83.23	97.37	95.75	69.73	93.84
C4	97.95	74.56	94.91	94.82	**54.39**	87.14	97.82	81.46	97.34	93.96	69.45	93.70
C5	98.04	74.10	93.66	94.05	52.54	84.93	97.84	80.89	96.93	94.18	69.55	92.91
C6	98.02	74.09	92.95	94.03	52.11	84.01	97.87	80.88	96.86	94.21	**69.84**	92.21
C7	98.25	77.35	**96.79**	61.76	42.78	63.03	**98.46**	85.39	**98.27**	78.84	66.28	75.32
C8	98.30	77.73	96.70	62.01	43.67	63.68	98.36	84.62	97.96	80.00	65.85	75.72
C9	98.29	77.89	96.54	61.81	43.66	63.73	98.33	84.53	97.98	79.79	65.99	75.91
C10	**98.36**	73.71	94.99	94.68	49.57	87.71	98.32	83.25	97.51	95.84	69.41	94.56
C11	98.37	73.79	94.55	94.66	49.76	**96.32**	98.26	83.21	97.29	**95.86**	69.46	**97.94**
C12	97.48	80.57	96.30	62.31	45.71	63.80	98.05	**86.27**	98.00	75.82	67.78	74.10
C13	97.74	80.65	96.15	62.09	46.49	64.38	98.04	85.50	97.75	78.26	67.08	74.68
C14	97.78	**80.71**	95.96	61.85	46.39	64.06	97.99	85.41	97.72	78.38	67.06	74.41
C15	98.28	77.73	96.49	61.87	43.11	63.93	98.36	84.62	98.07	79.77	66.63	75.50
C16	98.31	77.88	86.61	61.80	43.08	63.95	98.34	84.56	93.74	79.71	66.97	75.66
C17	98.00	73.39	93.52	94.08	53.31	85.61	97.86	80.86	96.99	94.30	69.26	92.72
C18	97.99	73.53	92.87	94.00	53.55	84.67	97.90	80.79	96.86	94.24	69.36	92.06
C19	97.72	80.63	95.98	61.92	46.06	64.02	98.05	85.49	97.81	77.91	67.79	74.36
C20	97.73	80.68	95.64	61.87	45.84	64.03	98.04	85.44	97.83	78.20	67.89	74.68
Avg.	98.07	76.61	94.82	78.21	48.04	75.66	98.14	83.68	97.39	86.83	68.23	84.56

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.
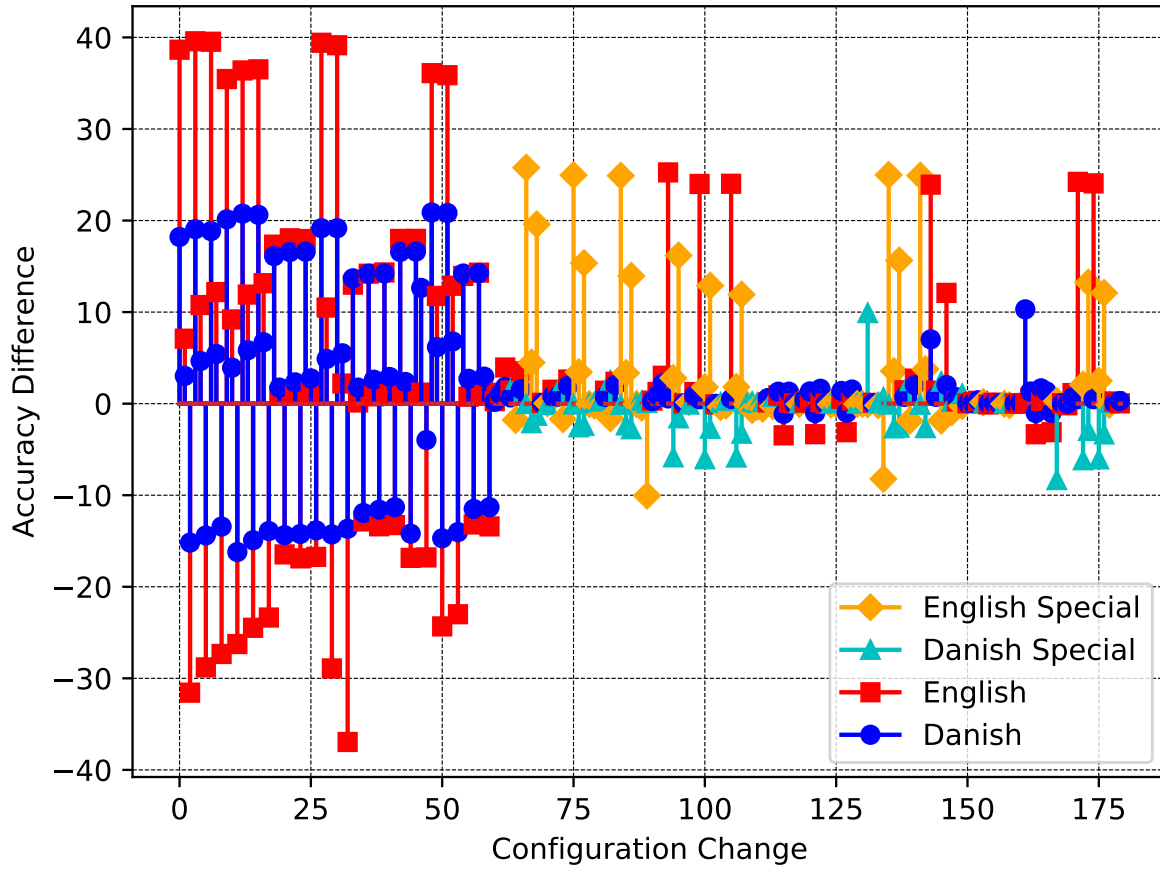
Fig. 2. Data points used in the Wilcoxon test. Configurations can be seen in Table IV. Each data point represents the percentage difference in accuracy between a pair of classification configurations. Data points marked with a blue circle represents the Danish data set and points marked with a red square 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.
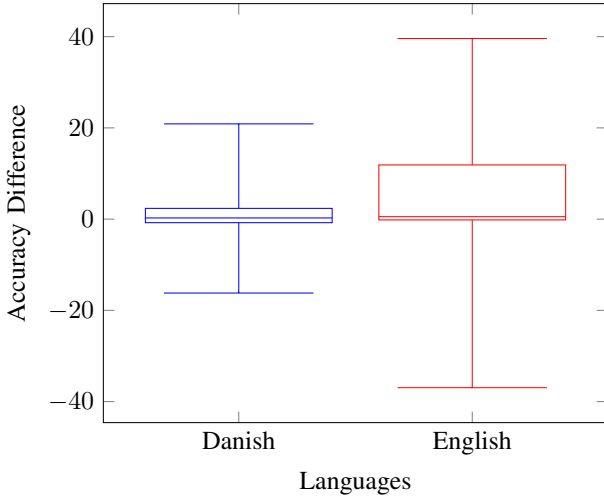- There are some differences between the Danish and

Fig. 3. Boxplot of data in Figure 2 which shows the accuracy difference using the quartiles; {Minimum, Lower Quartile, Median, Upper Quartile, Maximum} for the Danish and English data sets.

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 IV
THIS TABLE DESCRIBES THE X-AXIS IN FIGURE 2. EACH X-VALUE DESCRIBES A PAIR OF TEST CASE CONFIGURATIONS WITH ONLY ONE DIFFERENCE.

x	Config. Diff.	x	Config. Diff.	x	Config. Diff.	x	Config. Diff.	x	Config. Diff.	x	Config. Diff.
1	C1-SVM-NB	31	C11-SVM-NB	61	C1-C2-SVM	91	C4-C5-SVM	121	C8-C13-SVM	151	C13-C14-SVM
2	C1-SVM-RF	32	C11-SVM-RF	62	C1-C2-NB	92	C4-C5-NB	122	C8-C13-NB	152	C13-C14-NB
3	C1-NB-RF	33	C11-NB-RF	63	C1-C2-RF	93	C4-C5-RF	123	C8-C13-RF	153	C13-C14-RF
4	C2-SVM-NB	34	C12-SVM-NB	64	C1-C4-SVM	94	C4-C12-SVM	124	C8-C15-SVM	154	C13-C19-SVM
5	C2-SVM-RF	35	C12-SVM-RF	65	C1-C4-NB	95	C4-C12-NB	125	C8-C15-NB	155	C13-C19-NB
6	C2-NB-RF	36	C12-NB-RF	66	C1-C4-RF	96	C4-C12-RF	126	C8-C15-RF	156	C13-C19-RF
7	C3-SVM-NB	37	C13-SVM-NB	67	C1-C7-SVM	97	C5-C6-SVM	127	C9- C14-SVM	157	C14-C20-SVM
8	C3-SVM-RF	38	C13-SVM-RF	68	C1-C7-NB	98	C5-C6-NB	128	C9-C14-NB	158	C14-C20-NB
9	C3-NB-RF	39	C13-NB-RF	69	C1-C7-RF	99	C5-C6-RF	129	C9-C14-RF	159	C14-C20-RF
10	C4-SVM-NB	40	C14-SVM-NB	70	C2-C3-SVM	100	C5-C13-SVM	130	C9- C16-SVM	160	C15-C16-SVM
11	C4-SVM-RF	41	C14-SVM-RF	71	C2-C3-NB	101	C5-C13-NB	131	C9-C16-NB	161	C15-C16-NB
12	C4-NB-RF	42	C14-NB-RF	72	C2-C3-RF	102	C5-C13-RF	132	C9-C16-RF	162	C15-C16-RF
13	C5-SVM-NB	43	C15-SVM-NB	73	C2-C5-SVM	103	C5-C17-SVM	133	C10-C11-SVM	163	C15-C19-SVM
14	C5-SVM-RF	44	C15-SVM-RF	74	C2-C5-NB	104	C5-C17-NB	134	C10-C11-NB	164	C15-C19-NB
15	C5-NB-RF	45	C15-NB-RF	75	C2-C5-RF	105	C5-C17-RF	135	C10-C11-RF	165	C15-C19-RF
16	C6-SVM-NB	46	C16-SVM-NB	76	C2-C8-SVM	106	C6-C14-SVM	136	C10-C15-SVM	166	C16-C20-SVM
17	C6-SVM-RF	47	C16-SVM-RF	77	C2-C8-NB	107	C6-C14-NB	137	C10-C15-NB	167	C16-C20-NB
18	C6-NB-RF	48	C16-NB-RF	78	C2-C8-RF	108	C6-C14-RF	138	C10-C15-RF	168	C16-C20-RF
19	C7-SVM-NB	49	C17-SVM-NB	79	C2-C10-SVM	109	C6-C18-SVM	139	C10-C17-SVM	169	C17-C18-SVM
20	C7-SVM-RF	50	C17-SVM-RF	80	C2-C10-NB	110	C6-C18-NB	140	C10-C17-NB	170	C17-C18-NB
21	C7-NB-RF	51	C17-NB-RF	81	C2-C10-RF	111	C6-C18-RF	141	C10-C17-RF	171	C17-C18-RF
22	C8-SVM-NB	52	C18-SVM-NB	82	C3-C6-SVM	112	C7-C8-SVM	142	C11-C16-SVM	172	C17-C19-SVM
23	C8-SVM-RF	53	C18-SVM-RF	83	C3-C6-NB	113	C7-C8-NB	143	C11-C16-NB	173	C17-C19-NB
24	C8-NB-RF	54	C18-NB-RF	84	C3-C6-RF	114	C7-C8-RF	144	C11-C16-RF	174	C17-C19-RF
25	C9-SVM-NB	55	C19-SVM-NB	85	C3-C9-SVM	115	C7-C12-SVM	145	C11-C18-SVM	175	C18-C20-SVM
26	C9-SVM-RF	56	C19-SVM-RF	86	C3-C9-NB	116	C7-C12-NB	146	C11-C18-NB	176	C18-C20-NB
27	C9-NB-RF	57	C19-NB-RF	87	C3-C9-RF	117	C7-C12-RF	147	C11-C18-RF	177	C18-C20-RF
28	C10-SVM-NB	58	C20-SVM-NB	88	C3-C11-SVM	118	C8-C9-SVM	148	C12-C13-SVM	178	C19-C20-SVM
29	C10-SVM-RF	59	C20-SVM-RF	89	C3-C11-NB	119	C8-C9-NB	149	C12-C13-NB	179	C19-C20-NB
30	C10-NB-RF	60	C20-NB-RF	90	C3-C11-RF	120	C8-C9-RF	150	C12-C13-RF	180	C19-C20-RF

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 set[6], which is used for a semantic evaluation workshop.

REFERENCES

[1] M. V. Mäntylä, D. Graziotin, and M. Kuutila, "The evolution of sentiment analysis—a review of research topics, venues, and top cited papers," *Computer Science Review*, vol. 27, pp. 16 – 32, 2018. doi: 10.1016/j.cosrev.2017.10.002

[2] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American Scientist*, vol. 89, no. 4, pp. 344–350, 2001. doi: 10.1511/2001.4.344

[3] G. Angiani *et al.*, "A comparison between preprocessing techniques for sentiment analysis in twitter," in *KDWeb*, 2016. doi: 10.1007/978-3-319-67008-9_31

[4] W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth, "Harnessing twitter "big data" for automatic emotion identification," in *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, Sep. 2012. doi: 10.1109/SocialCom-PASSAT.2012.119 pp. 587–592.

[5] A. Balahur, "Sentiment analysis in social media texts," in *WASSA@NAACL-HLT*, 2013. doi: 10.1.1.310.4764

[6] B. Gokulakrishnan, P. Priyanthan, T. Ragavan, N. Prasath, and A. Perera, "Opinion mining and sentiment analysis on a twitter data stream," in *International Conference on Advances in ICT for Emerging Regions (ICTer2012)*, Dec 2012. doi: 10.1109/ICTer.2012.6423033 pp. 182–188.

[7] V. K. Jain, S. Kumar, and S. L. Fernandes, "Extraction of emotions from multilingual text using intelligent text processing and computational linguistics," *Journal of Computational Science*, vol. 21, pp. 316 – 326, 2017. doi: 10.1016/j.jocs.2017.01.010

[8] M. Asad, N. Afroz, L. Dey, R. P. D. Nath, and M. A. Azim, "Introducing active learning on text to emotion analyzer," in *2014 17th International Conference on Computer and Information Technology (ICCIT)*, Dec 2014. doi: 10.1109/ICCITechn.2014.7073079 pp. 35–40.

[9] J. R. Quinlan, "Induction of decision trees," *MACH. LEARN*, vol. 1, pp. 81–106, 1986. doi: 10.1007/BF00116251

[10] C. Cortes and V. Vapnik, "Support-vector networks," in *Machine Learning*, 1995. doi: 10.1007/BF00994018 pp. 273–297.

[11] J. C. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," in *Advances in Kernel Methods-Support Vector Learning*, 1999.

[12] T. Ho, "Random decision forests," in *Document Analysis and Recognition, International Conference on*, vol. 1, 09 1995. doi: 10.1109/ICDAR.1995.598994. ISBN 0-8186-7128-9 pp. 278 – 282 vol.1.

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

[13] G. F. Cooper and E. HERSKOVITS, "A bayesian method for the induction of probabilistic networks from data," in *MACHINE LEARNING*, 1992. doi: 10.1007/BF00994110 pp. 309–347.

[14] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.