UNIVERSITY OF MALTA
L-Università ta' Malta

INSTITUTE OF LINGUISTICS

M.Sc. Human Language Science and Technology

LIN5507: Social Media as Multimodal Text

Is Code-Switching related to Sentiment?

Xiaoyu Bai
Aina Garí Soler
Hoa Vu Trong

Malta, February 2017

Professor: Albert Gatt

Contents

1. Introduction

Code-switching (CS) has been defined as the use of two or more linguistic varieties in the same conversation or interaction (Scotton and Ury, 1977). It is a typical phenomenon in bilinguals that seems to take place universally, for any language pair. A large amount of CS in written form has also been observed in social media texts (Cárdenas-Claros and Isharyanti, 2009; Shafie and Nayan, 2013), and this is the kind of text the present study focuses on. Concretely, we use Twitter data.

There has been much linguistic work investigating the possible reasons for CS or factors that could motivate it. The following shows several reasons that are mentioned in the literature, among others:

- Conveying an exact meaning;
- Easing communication;
- Showing greater authority;
- Capturing attention;
- Creating a communication effect;
- Excluding someone;

  (Chen, 2003)

- Lack of register;
- Emphasizing a point;
- Showing identity with a group;
- Mood of the speaker.

  (Malik, 1994)

When describing the mood of the speaker factor, Malik (1994) claims that when the speaker is in a disturbed state of mind, such as in a state of tiredness or irritation, they might have difficulty in finding the appropriate words in a given language in which they are less proficient. This fits with the general intuition, derived from experience, that people switch language to argue or swear, and this is the starting point of our hypothesis. Our hypothesis is that emotion is one of the reasons behind CS, be it a positive or negative emotion. Therefore, we expect code-switching utterances (or tweets) to be emotional more often than monolingual ones.

In order to examine this, data annotated with sentiment is needed. We compare the proportion of tweets with different types of sentiment, and our study will reveal that, indeed, tweets with CS are significantly more often emotional than monolingual tweets, no matter the polarity of the emotion.

We also investigate whether the addition of a CS feature (in terms of presence or absence) can help a sentiment classifier to label tweets with sentiment more accurately, but it will be shown that our data do not seem to support this idea.

We begin by outlining the background and motivation for our investigation (Section 2), then proceed to describe our datasets (Section 3). Sections 4 and 5 detail two experiments carried out to examine our hypothesis, while Section 6 describes further research we perform on the basis of the insights gained from the previous sections. Finally, we discuss some limitations of our studies in Section 7 before Section 8 concludes the present report.

2. Background and Motivation

2.1. CS and Emotions in Spoken Language

The starting point of our investigation is rooted in the intuition that there exists a link between language and emotion. Indeed, the notion that bilingual or multilingual speakers' choice of language, including CS behaviour, is affected by emotions has been the subject of much research.

Among others, Besemeres (2004) offers a detailed study of narratives by bilingual authors, with an emphasis on the relationship between language and emotions. She finds out that certain emotional concepts and expressions are perceived as untranslatable, which prompts speakers to practice CS in order to express them. For instance, one of the authors in her studies states that "her 'internal monologue' [proceeds] for the most part in English, but [is] interrupted by Polish phrases often drawn from the emotional realm" (Besemeres, 2004:143).

Dewaele (2010: Ch.10) studies, inter alia, the types of conversation topics which incite higher frequencies of CS based on participants' self-report. The following graph, reproduced from p. 197, summarises his results:

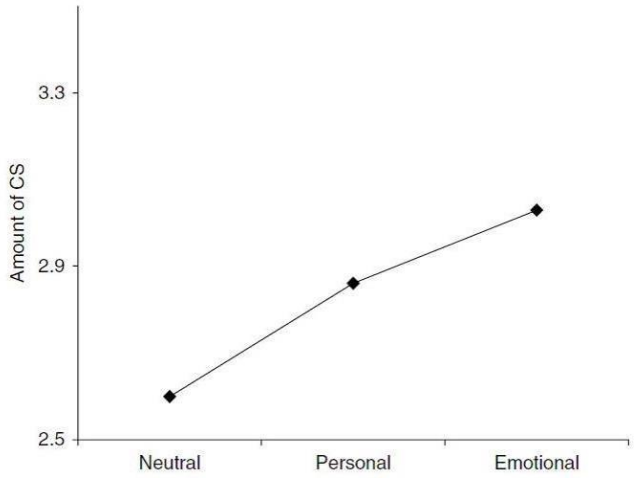

Figure 1: Mean frequency of self-reported CS according to nature of conversation topic, reproduced from Dewaele (2010:197)[1]

As illustrated in Figure 1, the author found a significant preference for self-reported CS in cases of *emotional* conversation topics.

Furthermore, in a previous study, Dewaele (2004) reveals that swear words are perceived as more forceful in one's native tongue (commonly referred to as L1) and that therefore his study participants in multilingual environments report switching to their L1 to express strong negative emotions. Similarly, the phrase "I love you" is equally regarded as strongest in their L1 by the majority of Dewaele (2008)'s study participants, while Pavlenko (2004) observes a tendency for parents in multilingual households to code-switch in order to express emotional phrases such as endearments in parent-child communication. It should be remarked that the perceived force of emotional expressions in one's L1 can favour CS both into and from the L1: The former case typically aims to lend more force to an emotional statement, whereas the latter allows speakers to make emotional utterances which would be perceived as too uncomfortably strong in the L1 (Pavlenko, 2007; Dewaele, 2010). For example, some of Dewaele (2004)'s study participants report swearing in non-L1 languages since it is perceived (by themselves) as less offensive. In either case, emotion acts as a significant factor in multilingual speakers' choice to practice CS.

[1] Dewaele (2010) does not specify the exact meaning of the values on the y-axis. Participants were asked to respond to questions concerning their CS frequency with "Never", "Rarely", "Sometimes", "Frequently" and "All the time" (p. 195).

2.2. Language Usage and Emotions in Social Media Text

Certainly, the aforesaid research centres on spoken language and conversations and as such could be argued to have no direct bearings on written data. Nonetheless, emotions may have the same influence on CS behaviour in Twitter text for the following reasons:

Thelwall (2009) remarks that scholars have long been investigating in how far internet language might be more similar to spoken language than to written language. While the verdict has been ambiguous and research places internet language either between written and spoken language or in a category of its own, there is little doubt that internet language bears characteristics of spoken language (Thelwall, 2009; Ko, 1996; Yates, 1996). To illustrate, Thelwall (2009)'s own research on MySpace comments shows that, in terms of the rank order of frequent words used, MySpace data resemble spoken rather than written data, characterised by pervasive usage of personal pronouns. Moreover, among the prevalent types of non-standard language usage in social media text are phonetic spelling (Thelwall, 2009), e.g. *Me traes 2 tacos d al pastor y 2 d fajita*[2], and expressive lengthening, e.g. *cooollllll* (Eisenstein, 2013). Both rely on an approximation to the *pronunciation* of the expressions in question. In addition, while social networks like Facebook, MySpace, Twitter, etc. are not intended for synchronous communication, i.e. communication in real-time, the common practice of commenting on posts, replying to comments and tweeting in reaction to previous tweets nonetheless create a dialogue-like character (see Figure 1 in Appendix for an example from Twitter). Thelwall (2009) terms the practice of writing on each other's walls or profile pages between social network friends as "public conversations".

Social media text further combines informality with emotion-laden content. Studies by Thelwall (2008) and Hinduja and Patchin (2008) expose substantial amounts of swearing in MySpace posts and comments, suggesting not only that social media postings tend to be highly informal, but that they also frequently seem to deal with contents which incite strong emotions. Thelwall (2009)

[2] "Bring me 2 'al pastor' tacos and 2 'fajita' tacos" (taken from our dataset), where *d* is used instead of the preposition *de* because the letter alone has the same sound as the preposition.

also identifies love (though not necessarily of a romantic nature) as a key theme in MySpace comments, as exemplified in posts such as “LIL SIS love you!!”.

Finally, various studies reveal CS to be frequent in social media text: Shafie and Nayan (2013) examine the language usage in the Facebook content of Malaysian University students and find that the students frequently practice CS both in their wall posts and comments; Cárdenas-Claros and Isharyanti (2009) show CS practice by Spanish and Indonesian non-native speakers of English in internet chats; and the first shared task on language identification in code-switched social media data, mainly from Twitter (Solorio et al., 2014), shows that CS on social media is large-scale enough to merit considerable attention from the NLP community.

In sum, social media text bears properties of spoken language, including spoken dialogues; it is largely informal, often deals with sentiment-laden content and contains considerable CS. As such, it shares many traits with the communicative situations mentioned in the previous section. Therefore, on the assumption that CS in spoken conversations is related to emotions and sentiment, there is ample reason to believe that the same relation might be reflected in social media data in general and Twitter data in particular. The present study aims at verifying this hypothesis.

3. Data used

This study requires Twitter data produced by users in a multilingual community known to practice CS. More specifically, we require a) tweets by such users that do not contain CS, which we refer to as ‘monolingual tweets’, and b) tweets by the same users that do contain CS, which are hereafter referred to as ‘CS tweets’. Sentiment analysis on both datasets will reveal if the CS tweets express more sentiment.

The data chosen for this task are retrieved from the first shared task on language identification in code-switched social media data (Solorio et al., 2014), made publicly available by the task

organisers[3]. Code-switched data from four language pairs were used in the shared task, from which we have chosen the data for the English-Spanish CS pair. Solorio et al. (2014) further provide the following information on the data collection process: In a preliminary step, they targeted tweets by users from California and Texas and extracted tweets which contain CS, i.e. both Spanish and English words. On the basis of this, they then identified 12 highly prolific writers of CS tweets and retrieved all of their available tweets, including both monolingual and CS ones. These tweets then constitute their final English-Spanish dataset. Of these, the present study uses the 11,400 tweets which have been designated as the training data in the shared task (Solorio et al., 2014).

As mentioned, the subsequent tasks in our study centre on sentiment analysis of both the monolingual and the CS tweets. Automatic sentiment analysers are unlikely to perform well on any kind of code-switched data, let alone Twitter data. However, Vilares et al. (2016) have extracted the CS tweets from Solorio et al. (2014)'s English-Spanish training dataset, i.e. the very data collection used for our study, and have provided manual sentiment labels for these tweets (3,062 in total). Their intention is to provide the research community with a set of gold-standard annotation against which to evaluate sentiment classification in multilingual social media data. The annotation has been carried out by three annotators and is available both in the scaled binary classification format of SentiStrength (Thelwall et al., 2010) and in the trinary scale format using the three categories "positive", "negative" and "neutral" (Vilares et al., 2016).

For the present study, instead of directly using English-Spanish CS tweets from Solorio et al. (2014), we use the dataset provided by Vilares et al. (2016)[4], which originates from the CS part of the aforesaid 11,400 tweets but is augmented with manual sentiment labelling as described above. The version in the trinary scale annotation format is used in the present study. With regard to monolingual tweets, as all the tokens in the 11,400 tweets from Solorio et al. (2014) are already equipped with labels for English and Spanish, it is possible to extract those tweets which contain only English tokens.

[3] http://emnlp2014.org/workshops/CodeSwitch/call.html
[4] http://www.grupolys.org/software/

Hence, in sum, our data collection process ultimately yields the following two datasets originating from Solorio et al. (2014)'s shared task data: a) a set of 3,062 English-Spanish CS tweets, enhanced with manual trinary scale sentiment annotation by Vilares et al. (2016), and b) a set of 5,488 English monolingual tweets produced by the same Twitter users[5], which has yet to undergo sentiment analysis.

It is appreciated that the two datasets have been produced by the same Twitter users. As the users involved are invariable, should the CS tweets be found to express more emotions, it cannot be argued that they do so because they originate from users who generally code-switch more than those who produced the monolingual tweets. At the same time, it should also be conceded that the entire dataset used in this study originates from only 12 Twitter users, which can be considered as undermining the study results. This point will be taken up again in Section 7.

4. Experiment 1

4.1. Method

To examine the hypothesis that CS is related to sentiment, we need to test the independence between these two variables. As we mentioned in Section 3, thanks to the publicly available datasets from Vilares et al. (2016) and Solorio et al. (2014), we already have 3,062 sentiment-annotated CS tweets and 5,488 unannotated English monolingual tweets. In this experiment, we employ a sentiment classifier to annotate the English monolingual tweets.

We need a classifier that can perform well on social media text, which is often short and noisy. The rule-based sentiment classifier Vader (Hutto and Gilbert, 2014) satisfies these requirements. Vader first solves the problem with a sentiment lexicon, which is an adaptation and enrichment of well-established sentiment lexicons from domains other than social media. They then

[5] Note that the two datasets do not add up to 11,400. This is the case because the original dataset from Solorio et al. (2014) also contains Spanish monolingual tweets as well as tweets containing tokens with an ambiguous language status (see the shared task annotation guideline for the English-Spanish data, available on http://emnlp2014.org/workshops/CodeSwitch/call.html). Those tweets are not considered in the present study.

add sentiment intensity ratings for each lexicon entry based on crowdsourcing data. Finally, on the basis of this lexicon, a rule-based classifier is created using heuristic rules. Vader performs very fast and, as a rule-based classifier, requires no training data.

Vader's output consists of individual scores for the categories "pos" (positive), "neg" (negative) and "neu" (neutral) as well as a compound score, as exemplified in the following:

```
VADER is smart, handsome, and funny.
{'neg': 0.0, 'neu': 0.254, 'pos': 0.746, 'compound': 0.8316}
```

In our experiment, we proceeded as follows: We regarded the individual, non-compound scores, assigning the entire tweet to the sentiment class "positive" if "pos" had the highest score among the three labels, to class "negative" if 'neg' had the highest score and, analogously, to class "neutral" if "neu" had the highest score. In order to be consistent with the sentiment labels used by Vilares et al. (2016), we further renamed the classes "pos", "neg" and "neu" as "P", "N" and "NONE", respectively. Table 1 below shows the classification result of Vader on the English monolingual tweets, juxtaposed with the manual annotation from Vilares et al. (2016) on the CS tweets:

	Monolingual tweets			CS tweets		
Label	P	N	NONE	P	N	NONE
Occurrence	478	278	4299	963	786	1313

Table 1: Results of trinary classification by Vader on the monolingual dataset and by Vilares et al. (2016) on the CS dataset

We also wished to examine the relation between sentiment and CS *regardless of the polarity of the sentiment*. Hence, we conflated the classes "P" and "N" to create the class "Sentiment" in opposition to the class "Non-Sentiment". The latter simply corresponds to the class "NONE" in the trinary classification. Table 2 below shows the classification results in this binary setting:

	Monolingual tweets		CS tweets	
Label	Sentiment	Non-sentiment	Sentiment	Non-sentiment
Occurrence	756	4299	1749	1313

Table 2: Results of binary classification by Vader on the monolingual dataset and by Vilares et al. (2016) on the CS dataset

Based on these data, we investigated the potential dependency between CS and sentiment using the Pearson's Chi-Squared test of independence. In our case, the categorical variable "CS" can take on the values presence ("CS tweets") or absence ("Monolingual tweets") of CS; while the categorical variable "sentiment" can have the values "P", "N" and "NONE" in the 3-way classification setting and presence ("Sentiment") and absence ("Non-sentiment") of sentiment in the 2-way classification setting.

4.2. Results and Discussion

The Chi-Squared test was conducted individually for both classification settings. In both cases, the result shows a statistically extremely significant dependency between CS and sentiment ($p < 0.00001$, $\chi^2 = 1601.76$ in the 3-label setting and $p < 0.00001$, $\chi^2 = 1661.21$ in the 2-label setting). This seems to prove our hypothesis is true, given this dataset. However, upon further investigation, we spot some minor issues which lead us to question the validity of these numbers:

First, output from Vader seems extremely biased toward the category "neutral". The reason is that Vader uses a small amount of high-confident manual rules which make the classifier overlook many instances of sentiment in tweets. For example, "Watched @username for a 2nd time today! @username & @username are too damn hilarious! Entire room was gone laughing the whole time! xD" is clearly positive but marked as neutral by Vader. Second, annotation on the CS dataset provided by Vilares et al. (2016), on the other hand, seems to us to be biased toward the categories "P" and "N", i.e. those expressing the presence of sentiment. For instance, the tweet "@username ahhhhh, sorry entendi mal, ya relei lo que dices. enviamelo por fa por correo :)"[6], in our judgment, does not express any emotion but is labelled as positive by Vilares et al. (2016). The root of the problem here is that the two sides of our data, monolingual and CS, are not sentiment-labelled with the same method. The fact that the former is automatically classified by Vader and the latter annotated by hand appears to be a serious drawback.

[6] ahhhhh, sorry I understood it wrongly, I already re-read what you say, please send it to me by (e-)mail :)

Moreover, upon closer inspection, it was revealed that the alleged CS tweets (in the sense of containing both English and Spanish tokens) by one of the 12 users do not in fact constitute CS. The user frequently teaches English on Twitter by posting English phrases and their Spanish translations, along with an English example sentence. To illustrate, an instance is the following tweet: "Work my butt off: Trabajar mucho. I work my butt off so that my children can have everything they need"[7]. Clearly, this tweet merely *mentions* English and Spanish tokens on a meta-level and is therefore not a case of natural language *usage*. For the purpose of training and testing language identification algorithms, such tweets could arguably be included, yet for the purpose of investigating sentiment in CS tweets, they are clearly ill-suited.

To solve these issues, we carry out a second experiment along the same lines, using, however, a small sample of data for which we provide our own manual annotation. The sections to come outline this follow-up investigation.

5. Experiment 2

5.1. Method

As a first step of this second experiment, we removed from our dataset all the tweets by the above-mentioned user. After the removal, our complete dataset comprises 2237 CS and 4884 English monolingual tweets. From these, 500 CS and 500 monolingual tweets were randomly extracted to be manually annotated for sentiment.

The manual annotation of these 1000 tweets was carried out by all three authors of the present report. In terms of annotation method, we were guided by Wiebe et al. (2005)'s instructions for annotating emotions in language. It should be remarked that the task nonetheless posed a significant

[7] It should be remarked that this very tweet has been annotated as "P" in Vilares et al. (2016), which strengthened our reluctance to rely entirely on their annotation. To our mind, even if this tweet did occur in a context of natural language usage, it could not be regarded as expressing positive sentiment.

challenge since Wiebe et al. (2005: 31) explicitly states that no formal criteria can be given for guidance and that annotators are to rely on their "human knowledge and intuition".

To obtain a similar format to the manual annotation from Vilares et al. (2016), we annotated a given tweet as "positive (P)" if it clearly expresses a positive emotion or opinion, "negative (N)" if it clearly expresses a negative emotion or opinion and "neutral (NONE)" if it does not express any obvious or apparent emotion. Furthermore, we added a fourth category, "mixed (M)", which applies a) to tweets which clearly express both positive and negative emotions in a comparable amount and b) to tweets which express emotions that cannot be distinctly classified as either positive or negative. The addition of the category M was motivated by tweets like "Not sleepy yet!! :o", where the two exclamation marks and the emoticon make the category NONE seem inappropriate, but where it is also impossible to clearly classify the tweet as P or N. The idea here is that while tweets of the category M cannot be classified as P or N, they can be unambiguously included in the category "Sentiment". Table 3 below presents each label of our 4-way classification with an example tweet from our dataset:

Label	**Tweet text**
P	Tomorrow shall be good 😍
N	Damn I fucked up
M	DONE!!!!!!!!!
NONE	Time for work

Table 3: Examples of our annotation for the sentiment labels "P", "N", "M" and "NONE".

Furthermore, we took note of the following: Generally, we labelled the sentiment expressed by each tweet on its own. However, where the tweet was clearly a response to a previous tweet and difficult to understand on its own, we also took into consideration its preceding context. Moreover, as much as possible, we labelled the tweets strictly according to the emotions they *express*, based on our human knowledge, but refrained from making assumptions regarding the user's likely mood when composing the tweet. To illustrate, while the tweet "Florida with the family #BestMemoriesOf2013", particularly

with the hashtag *#BestMemoriesOf2013*, allows us to surmise that its writer was likely in a positive mood at the time of its composition, the tweet itself does not express any positive sentiment and therefore would be labelled as NONE.

We tested the inter-annotator agreement between the three annotators on 30 English monolingual tweets, using the Fleiss' Kappa value. At 0.589, the agreement is "moderate" according to Landis and Koch (1977) and only marginally below the category "substantial agreement" (for kappa values between 0.61 - 0.80). Although consistency in annotation was attempted, due to the subjective nature of the task and the absence of clearly defined formal guidelines, the inter-annotator agreement is modest, yet it is passable. The result of our manual annotation is summarised in Table 4 below:

	Monolingual tweets				CS tweets			
Label	P	N	M	NONE	P	N	M	NONE
Occurrence	116	72	65	247	145	95	69	191

Table 4: Manual annotation results of 500 monolingual and 500 CS tweets in the 4-label setting

Once again, we also represent our annotation results in terms of the categories "Sentiment" and "Non-sentiment", where the former corresponds to the sum of the categories "P", "N" and "M" and the latter to NONE. This is shown in Table 5 below:

	Monolingual tweets		CS tweets	
Label	Sentiment	Non-sentiment	Sentiment	Non-sentiment
Occurrence	253	247	309	191

Table 5: Manual annotation results of 500 monolingual and 500 CS tweets in the 2-label setting

As in the previous experiment, we conducted the Pearson's Chi-Squared test of independence to investigate the association between the emotions in tweets and the presence /absence of CS behaviour, based on our manual annotation. We again did so in two settings: In the first, the categorical variable "CS" has the values "Monolingual" and "CS" and the variable "sentiment" the values "Positive (P)", "Negative (N)" and "Neutral (NONE)". As previously suggested, in this setting, we excluded the category "Mixed (M)", of which we can only be certain that it signals emotions, but not whether the emotion in question should be classified as positive or negative. Moreover, doing so, we obtained the

same three emotion categories which we used in the previous experiment, based on the annotation by Vilares et al. (2016). In the second setting, we again used the presence/absence of emotions, i.e. "Sentiment" and "Non-Sentiment", on the one hand and the presence/absence of CS, i.e. "Monolingual tweets" and "CS tweets", on the other.

In addition to the above, based on our manual annotation, we added a further two-by-two Chi-Squared test, using the values P and N for the variable *Emotion*. This was prompted by the intuition mentioned in the introduction as well as Dewaele (2004)'s study on swear words, which seems to highlight speakers' decision to code-switch in order to express *negative* emotions. Therefore, among the tweets labelled with P and N, we examined if any association exists between the positivity/negativity of sentiment in tweets and CS behaviour.

5.2. Results and Discussion

Pearson's Chi-Square in the first setting (using "P", "N" and "NONE") shows a statistically significant, medium to strong association between the variables "sentiment" and "CS" ($p < 0.01$, $\chi^2 = 13.532$; Cramer's $V = 0.46$), with CS tweets expressing more emotions overall. To illustrate, with respect to the label "NONE", our data count 247 occurrences in the monolingual tweets and 191 in the CS tweets, compared to their respective expected counts of 220.01 and 217.99[8].

Similarly, in the second setting, the presence/absence of sentiment shows a statistically highly significant and medium association with the presence/absence of CS ($p < 0.001$, $\chi^2 = 12.74$; Cramer's $V = 0.40$). In particular, since this test makes use of all of our labelled tweets, among which exactly 500 are monolingual and 500 contain CS, the counts in Table 5 are sufficient to reveal that CS tweets express sentiment more frequently than monolingual tweets do. The third Chi-squared test, intended to test for association between the polarity of emotions expressed in tweets and CS, does not yield a statistically significant result ($p = 0.787$).

[8] Expected counts in a Chi-Squared test show the counts one would expect for each cell of the contingency table *if no association at all* existed between the two categorical variables in question.

From these numbers the following can be concluded: There exists an association between emotions expressed in twitter data and CS behaviour by bilingual or multilingual twitter users. More precisely, tweets containing CS convey sentiment with a significantly higher frequency than monolingual tweets do. This confirms the hypothesis the present study intended to investigate. The tentative conjecture that more negative than positive emotions are expressed in CS tweets is not borne out as we see no attested association between the positivity/negativity of emotions in tweets and the presence/absence of CS.

6. Further Studies Carried Out

6.1. Training a Sentiment Classifier to Extend our Annotation

Our study has shown that CS texts on Twitter express emotions more often than monolingual (or non-CS) texts, based on a sample of 1000 manually sentiment-labelled tweets. From here, we identified two avenues for further work and examined both:

The first possible continuation of the present study is to train a sentiment classifier on the basis of our 1000 monolingual and CS tweets with manual sentiment-annotation. More specifically, the idea is to train a sentiment classification algorithm separately on the 500 monolingual and 500 CS tweets. The resulting classifiers would then be applied to the rest of our original set of monolingual tweets and CS tweets, respectively[9]. We would then obtain uniformly classified sentiment labels for our entire dataset in the sense that both the monolingual and CS data would be classified using the same classification algorithm and comparably labelled training data, which is in contrast to our previous experiment based on the annotation by Vilares et al. (2016) and Vader output (see the points raised in Section 4.2).

[9] Recall that prior to extracting the 1000 tweets for manual annotation, our original dataset comprises 4884 monolingual and 2237 CS tweets (see Section 5.1).

Using the NLTK toolkit (Bird et al., 2009), we implemented a Maximum Entropy classifier with the following features based on the work of Mansour (2015):

- Bag of words;
- # of times emoticons from a pre-specified set occur in the tweet (the set was created from emoticons in the data);
- Proportion of capital letters;
- # of exclamations;
- # of positive and negative words according to SentiWordNet (Baccianella, 2010).

However, as shown in the following, the classifier's performance was revealed to be too low to be useful: First, we designed the task as a classification task into the output classes P, N, M and NONE. The baseline, defined as taking the most frequent class encountered during training, is 0.494 for the monolingual and 0.382 for the CS tweets (with NONE being the most frequent class in either case). For both sides of our data, i.e. monolingual and CS, we trained on 350 randomly selected tweets out of our 500 labelled tweets and tested on the remaining 150 tweets. For either side, we executed 100 runs, re-selecting in each run the training dataset at random, and took the average classification accuracy. We summarise our results in Table 6. The classification result was 0.534 for monolingual and 0.536 for CS tweets. While better than the baseline, these numbers evidently show extremely low classification performance.

In a second step, we changed the task to binary classification, with the classes "Sentiment" and "Non-sentiment" as output. The baseline in this case is classifying all tweets as "Sentiment", which would yield a classification accuracy of 0.506 for the monolingual and 0.618 for the CS datasets. Once again, for both sides of the data, we divided the labelled data into 350 tweets for training and 150 tweets for testing. The classification accuracies we recorded based on the average of 100 runs are 0.622 for the monolingual data and 0.627 for the CS data, which is superior to the baseline, although only marginally so in the case of the CS data.

Clearly, if we are to reliably classify the *unlabelled* tweets in our dataset in order to test for associations between emotions and CS behaviour in our complete dataset, we would require a better performance from the classifier. Even if we were to use all 500 labelled tweets for training (for either

side of the data) and did not hold out any for testing the classification performance, the improvement would likely not be substantial. We believe that 350 – 500 tweets are insufficient to train a reliable sentiment classifier.

Train-test size	Labels	Dataset	Baseline accuracy	Accuracy achieved
350-150	P, N, M, NONE	Monolingual	0.494	0.534
		CS	0.382	0.536
	SENT, NONE	Monolingual	0.506	0.622
		CS	0.618	0.627

Table 6: Classification accuracies of the first experiment on all settings

6.2. CS as a Feature for Sentiment Classification

The second possibility for further research using the insight gained from our studies is as follows: Since CS tweets express emotions more often than monolingual tweets do, a given CS tweet should have a higher probability of expressing sentiment than a monolingual one. Therefore, it is conceivable that the presence/absence of CS in a given tweet could be a useful feature for a classifier in determining whether or not the tweet is sentiment-laden.

Pursuing this line of thought, we mixed and shuffled the 1000 CS and monolingual tweets with manual annotation. We then trained and tested the same binary Maximum Entropy classifier (classifying into "Sentiment" and "Non-sentiment") with, however, the following alteration: In the first setting we used the same features as previously described (bag of words, emoticons, capital letters, exclamation marks, information from SentiWordNet); in the second, we added the feature "CS" with the possible values "True" (for presence of CS) and "False" (for absence of CS). The 1000 tweets were divided into a training set of 700 randomly selected tweets and a test set consisting of the remaining 300 tweets. As in the previously described classification tasks, in each of the two settings (with and without the CS feature) we executed 100 runs and took the average accuracy. Table 7 below

summarises our results. For the setting *excluding* the CS feature, we recorded a classification accuracy of 0.671. In the setting *including* the CS feature, the performance was identical. Thus, contrary to our expectations, the performance in the second setting is not superior to that in the first. We subsequently increased the size of the training data, now training on 900 and testing on 100 tweets. In this configuration, the average accuracy of 100 runs was 0.689 *excluding* the CS feature and 0.681 *including* it. Once again, the inclusion of CS as a feature for the binary classifier did not entail an improved performance (we in fact see a minor decrease in accuracy).

These results are surprising, given that our previous tests of association have distinctly demonstrated a strong association between sentiment in tweets and CS behaviour, with CS tweets expressing emotions more often than monolingual tweets. Yet, the hypothesis that CS could be used as a helpful feature in sentiment classification is not confirmed. Nonetheless, at 1000 tweets, we suspect that our training dataset might be too small to reveal any possible benefit of a selected feature. Therefore, we do not completely reject the idea of using CS as a classification feature but suggest testing it further in future studies based on more labelled data.

Dataset	Train-test size	Labels	CS feature	Accuracy achieved
Monolingual + CS	700-300	SENT, NONE	no	0.67
			yes	0.671
	900-100		no	0.689
			yes	0.681

Table 7: Classification accuracies of the CS-feature experiment on all settings

7. Limitations of our Study

Here we discuss the shortcomings and weak points of our study, and how, where possible, future studies along the same line could avoid them:

Representativeness of our sample. As has already been said, our sample of tweets came from only 12 twitter accounts, a number that was reduced to 11 when we removed the user who was teaching English and used Spanish translations, which we did not consider to be real CS. Certainly, it could be argued that such a small sample is not representative of the CS community in general. In particular, our data have been supplied by twitter users from a specific community; therefore, our results first and foremost hold for this very group. Future research on a larger scale is needed to investigate in how far our conclusions can be generalised to CS communities at large. The limitation in the number of users comes from the unavailability of a more representative dataset and increasing it was beyond the scope of this study. Solorio et al. (2014) provided CS data for other languages but these had not been manually annotated with sentiment like the Spanish-English CS data.

Low inter-annotator agreement. The reported Fleiss' Kappa value, 0.589, is not a very high one and is below other values found in the literature (Takala et al., 2014). It should be taken into account that we only used 30 tweets to calculate the agreement; a bigger subset could have probably been a better estimate of the agreement. In any case, devising sentiment annotation guidelines for this task could have been helpful, but this would have required having already some previous experience in sentiment annotation of tweets. We take this as possible further work that could improve the agreement and, potentially, also the classifier's performance, since it has been argued that the inter-annotator agreement rate constitutes an upper bound for a sentiment classifier's accuracy (Mozetič et al., 2016).

Shortage of manually annotated data. As said in Section 6.1 the low amount of manually annotated tweets is probably the main reason behind our classifier's relatively low performance. It is not clear how much data would be necessary to achieve a significant improvement; however, given sufficient time and resources, we do not rule out repeating the experiment with more sentiment-labelled tweets in the future.

8. Conclusion

In this study, we presented our work on the relation between code-switching and emotion. We first compared the emotion in code-switched and monolingual tweets that had been labelled with sentiment manually and automatically, respectively. After this first comparison, however, we realized that having two different systems for the annotation of the two groups could be adding some bias. Moreover, we found one user whose data we did not consider cases of CS. Therefore, a second experiment was carried out in which we performed manual annotation of a subset of both CS and monolingual data, so as to ensure their comparability. Finally, a sentiment classifier was created and trained on our manually annotated data with two different goals: labelling the rest of the data with sentiment, and exploring the usefulness of a CS feature in such a classifier.

From the statistical analysis of the data that we annotated, we can conclude that CS utterances are more often emotional than monolingual ones, at least in the given circumstances (this specific language pair, Twitter data, these users), which gives support to our hypothesis that sentiment can motivate CS. Moreover, we show that, contrary to intuition, it is not the case that negative emotions are more often expressed than positive ones in CS data in comparison to monolingual data; rather, CS seems to be related to general excitement or fervor. However, all these results have to be taken carefully, considering the mentioned limitations of our study, viz. the low amount of Twitter users captured in our dataset and our modest inter-annotator agreement.

In relation to the classifier, it was not possible to use it to label the rest of the data due to its low accuracy, which we think is mainly caused by an insufficient amount of data. The addition of a CS feature did not improve the performance of the classifier. However, as already mentioned, we believe that due to the limited size of our data, we cannot pass any conclusive verdict on the usefulness of a CS feature. Future studies based on more labelled data should be conducted to explore efficient ways of exploiting the association between CS in social media text and sentiment.

9. References

Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *LREC* (Vol. 10, pp. 2200-2204).

Besemeres, M. (2004). Different languages, different emotions? Perspectives from autobiographical literature. *Journal of Multilingual and Multicultural Development*, *25*(2-3), 140-158.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Cárdenas-Claros, M. S., & Isharyanti, N. (2009). Code-switching and code-mixing in internet chatting: Between'yes," ya,'and'si'-a case study. *The Jalt Call Journal*, *5*(3), 67-78.

Cheng, K. K. Y. (2003). Code-switching for a purpose: Focus on pre-school Malaysian children. *Multilingua*, *22*(1), 59-78.

Dewaele, J. M. (2004). The emotional force of swearwords and taboo words in the speech of multilinguals. *Journal of multilingual and multicultural development*, *25*(2-3), 204-222.

Dewaele, J. M. (2008). The emotional weight of I love you in multilinguals' languages. *Journal of Pragmatics*, *40*(10), 1753-1780.

Dewaele, J. (2010). *Emotions in multiple languages*. Springer.

Eisenstein, J. (2013, June). What to do about bad language on the internet. In *HLT-NAACL* (pp. 359-369).

Hinduja, S., & Patchin, J. W. (2008). Personal information of adolescents on the Internet: A quantitative content analysis of MySpace. *Journal of adolescence*, *31*(1), 125-146.

Hutto, C. J., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Ko, K. K. (1996). Structural Characteristics of Computer-Mediated Language: A Comparative Analysis of InterChange Discourse. *Electronic Journal of Communication/La revue électronique de communication*, *6*(3), n3.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

Malik, L. (1994). *Sociolinguistics: A Study of Codeswitching*. New Delhi: Anmol

Mansour R., Hady M.F.A., Hosam E., Amr H., Ashour A. (2015) Feature Selection for Twitter Sentiment Analysis: An Experimental Study. In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2015. Lecture Notes in Computer Science, vol 9042. Springer, Cham

Mozetič, I., Grčar, M., & Smailović, J. (2016). Multilingual Twitter sentiment classification: The role of human annotators. *PloS one*, *11*(5), e0155036.

Pavlenko, A. (2004). 'Stop Doing That, Ia Komu Skazala!': Language Choice and Emotions in Parent—Child Communication. *Journal of multilingual and multicultural development*, *25*(2-3), 179-203.

Pavlenko, A. (2007). *Emotions and multilingualism*. Cambridge University Press. Scotton, C. M., & Ury, W. (1977). Bilingual strategies: The social functions of code-switching. *International Journal of the sociology of language*, *1977*(13), 5-20.

Shafie, L. A., & Nayan, S. (2013). Languages, code-switching practice and primary functions of Facebook among university students. *Studies in English Language Teaching*, *1*(1), 187.

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Gohneim, M., ... & Fung, P. (2014, October). Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching* (pp. 62-72).

Takala, P., Malo, P., Sinha, A., & Ahlgren, O. (2014). Gold-standard for Topic-specific Sentiment Analysis of Economic Texts. In *LREC* (Vol. 2014, pp. 2152-2157).

Thelwall, M. (2008). Fk yea I swear: cursing and gender in MySpace. *Corpora*, *3*(1), 83-107.

Thelwall, M. (2009). MySpace comments. *Online Information Review*, *33*(1), 58-76.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, *61*(12), 2544-2558.

Vilares, D., Alonso, M. A., & Gómez-Rodríguez, C. (2016). En-es-cs: An english-spanish code-switching twitter corpus for multilingual sentiment analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 4149-4153).

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, *39*(2), 165-210.

Yates, S. J. (1996). Oral and written linguistic aspects of computer conferencing. *Pragmatics and beyond New Series*, 29-46.

10. Appendix

Figure 1: Anonymised example of tweets with dialogue-like character.[10]

[10] Haha si tu y yo naci ayer = Haha yes and I was born yesterday
Ya tengo tiempo k no voy = I haven't gone there for a long time