Analyzing and Detecting Emotion Spams on Twitter

Qian Yang* **Ziyao Chen*** **Yu Zhao*** **Kyung Eun (Alex) Hur***

*School of Library, Archival and Information Studies
University of British Columbia

Abstract

Studying emotion-charged tweets on Twitter provides valuable information for various research areas. However, researchers face difficulties in studying such emotion tweets due to the presence of emotion spams—tweets that do not use their emotion hashtags for expressing emotions. In this study, we collected data set of 4,000 human-labelled tweets and conducted an in-depth analysis of emotion spams within the data set. The study identified seven categories of the emotion spams using ground-up approach, distributions of each categories, and important features of the emotion spams. The study also explores the emotion consistency between the tweet's text and the image attached. A machine learning model with 88.13% accuracy was then developed to identify the major emotion spams.

1 Introduction

In the digital era, people are increasingly sharing their emotion through social networking sites, such as Twitter and Facebook (Bazarova et al., 2015). Twitter in particular expresses emotions less formally than other social networking platforms as tweets are restricted to 140 characters and often written on mobile devices(Roberts et al., 2012). Due to the tight word limit, users often utilize the hashtag function where they can assign a main topic or theme of the text. Among 500 million tweets being tweeted every day on average, a number of them is used to express emotion using emotion hashtags (e.g. #sohappy and #disappointed). These emotion tweets provide a vast amount of information on research in various areas, such as studying human communication, analyzing product or event reviews, or understanding emotional responses toward controversial issues and international affairs (Roberts et al., 2012).

Despite the benefits of studying emotion tweets, researchers face difficulties in analyzing emotional tweets because more than 3% of them are spam on Twitter. A research by Bilge et al. (2009) found that 45% of users on a social networking site readily click on links posted by any friend in their friend lists' accounts, even if they may not know that person in real life. This fact attracts spammers to use Twitter as a tool to send unsolicited messages to legitimate users, post malicious links, and hijack trending topics (Mccord and Chuah, 2011). Spammers are prevalent in various tweet topics including emotion tweets where they use emotion spams, which are tweets that have promotional nature within the context of emotional tweets.

In order to better understand these emotion spams, and help researchers in conducting emotion analysis on Twitter, we conducted in-depth analysis on tweets with emotion hashtags and implemented a spam detection system. Section 2 of our paper discusses the related works of our study through literature review. Section 3, describes the methods we used and procedures we took in collecting and analyzing the data. Section 4 presents in-depth analysis of the corpus and related findings. Section 5 introduces the procedures of classifying the tweets using machine learning. Lastly, Section 6 points out the limitations, explores the future work, and summarizes the conclusion of our study.

2 Literature Review

2.1 Labeling Emotions of Tweets

There are two main approaches to label emotions of tweets (Kunneman et al., 2014). The first is to annotate the data by human experts (Roberts et al., 2012; Yan et al., 2016; Yan and Turtle, 2016). The second approach is to use the emotion hashtags that Twitter users added to a tweet (Mohammad, 2012; Mohammad and Kiritchenko, 2015; Kunneman et al., 2014; Qadir and Riloff, 2014).

Roberts et al. (2012) used human labeling approach to get accurate annotations of the emotions in tweets. They annotated seven emotions based on Ekman's (1992) six basic emotions—anger, disgust, fear, joy, love, and surprise—and love. From a total of 7,000 tweets, 57% were found to have no emotion. Among the tweets containing emotions, the most common emotions were *disgust* (16.4%) and *joy* (12.8%), followed by *anger* (10.4%), *love* (9.2%), and *sadness* (8.8%). Both *surprise* (5.8%) and *fear* (4.0%) were relatively rare. An analysis of the linguistic style of each emotion was also conducted using General Inquirer. Yan et al. (2016) developed a large emotion tweet corpus, EmoTweet-28, which contains over 15,000 tweets that were randomly sampling tweets and applying human labeling. Their annotation of each tweet includes valence (positive, negative, neutral), arousal (level of emotions), and 28 types of emotions. This is so far the most comprehensive work in classifying emotion of tweets through human labeling. The EmoTweet-28 emotion tweet corpus was then used to build a multiclass emotion classifier (Yan and Turtle, 2016).

Although human labeling may be a good way to obtain accurate labeling of emotions, it is time-consuming and can thus limit the size of the labeled data. Moreover, Wang et al. (2012) argued that the emotion labels assigned by people other than the author is not as reliable as the annotations made by the author itself. The hashtags on Twitter can be arguably viewed as the author's own annotation of its text and serve conversational purpose (Huang et al., 2010). To validate the consistency between emotion hashtags and emotion labels assigned by human, Mohammad (2012) presented several experiments and obtained consistent result. Therefore, many researchers use the hashtagging system on Twitter to obtain emotion-annotated data based on the *distant supervision* approach (Mohammad, 2012; De Choudhury et al., 2012; Purver and Battersby, 2012; Kunneman et al., 2014; Qadir and Riloff, 2014; Mohammad and Kiritchenko, 2015). *Distant supervision*, is an approach that leverages uncontrolled labeling to obtain large amount of training data (Mintz et al., 2009).

Yet, not all emotion hashtags are equally suitable to be used to label the emotion of tweets. Mohammad (2012) argues that emotion hashtags are included in tweets by users in two different ways: 1) *strengthen* the emotion already present in the tweet and 2) *add* emotion to an otherwise emotionally neutral message. It was argued that the examples of the second function will not provide proper training data for the detection of the emotion linked to the hashtag (Kunneman et al., 2014). Moreover, studies by Mccord and Chuah (2011) and Antonakaki et al. (2016) have identified the existence of large amount of spams on Twitter that abuse hashtags, which might include certain emotion hashtags. Huang et al. (2010) employed statistical measures to study hashtag usage patterns. They plotted time series data against the daily usage of a specific hashtag onto a scatterplot and discovered that standard deviation and kurtosis can be used to identify the nature of a hashtag—hashtags that have small standard deviations are usually referred to as micro-memes and kurtosis can be used to differentiate between micro-memes, recurring tags, and spams. However, such findings might not apply to a subset of emotion hashtags, and therefore can not be used directly to filter out unnecessary tweets. Therefore, a deeper analysis of the tweets that do not use their emotion hashtags for expressing emotions—*emotion spam*—might inspire ways to improve the quality of emotion labeled data obtained through distant supervision.

2.2 Tweets Classification

As Hayden White stated, "the beginning of all understanding is classification." Classifying emotion spams is the starting point to understand them. However, no prior work has been done to classify emotion spams in particular, but some work has been done for classifying tweets in general. Sriram et al. (2010) proposed an approach to classify tweets with a focus on user intention on Twitter by using author information and features extracted from the tweets. They used an available implementation of Naïve Bayes classifier to classify tweets into the following categories: news, events, opinions, deals, and private messages. Even though their experimental results showed that their approach outperformed the traditional Bag-Of-Words (BOW) strategy, the features of each tweet category need further scrutiny. For example, private messages were captured by having

'@username' at the beginning of tweets in their study. However, there are tweets that have '@username' at other positions in the tweet, but should be categorized as private message. There is also the possibility for users to abuse the '@username' for purpose other than private messaging. The authors also admitted that the noise in data might degrade the performance of their proposed approach. Naaman et al. (2010) applied human coding and quantitative analysis to gain a deeper understanding of the activity of individuals on Twitter. In their study, 350 users with public updates and are neither organizations or marketers were included in their study. In particular, a content-based categorization of the type of messages posted by Twitter users was created through a ground-up approach, which organized 3379 tweets into eight categories: information sharing, self promotion, opinions, statements, me now, questions, presence maintenance and anecdote. The categories emerged from the tweets revealed more valuable information about the content of the tweets than pre-defined categories.

2.3 Spam Detection

Even though the *emotion spam*, which our study focuses on, is different from the commonly acknowledged *spam* on Twitter, the works on detecting spam can provide a worthy reference. Some studies of spam detection on Twitter focus on user-based approaches (Wang, 2010a; Benevenuto et al., 2010). This method examines features related to user information. For example, Benevenuto et al. (2010) examined the fraction of followers per followees, age of the user account, and the number of tweets received from followees. Their study was able to reach an accuracy rate of 87.6% in the overall detection of spammers. There are also studies that solely looked at the user names to identify spam accounts. One example a study by Freeman (2013) where he was able to reach an accuracy of 85% and a false positive rate of 3.3%.

Studies by Antonakaki et al. (2016), Chen et al. (2014), and Benevenuto et al. (2010) focused on content-based approaches. Benevenuto et al. (2010) proposed 10 features to help detect spam tweets including the numbers of words from a list of spam words, hashtags per words, URLs per words, words, numeric characters in the text, characters that are numbers, URLs, hashtags, mentions, and times the tweet has been replied. By using these features, the classifier was able to correctly identify 78.5% of spam and 92.5% of the non-spam tweets with overall accuracy of 87.2%. Another example is in the Martinez-Romo and Araujo (2013)'s study. They compared the language model used between normal and suspicious tweets that are related to trending topics on Twitter and calculated the divergence between these two models. By using this language model strategy alone, the classifier was able to reach 91.6% accuracy and 88.7% for true positives. Moreover, by combining the features extracted from the language model and some of the features proposed by Benevenuto et al. (2010), the rate increased to 92.2% for accuracy and 89.3% for true positives.

URL is another aspect that has gained attention from researchers in the domain of spam detection. Various features extracted from the URL of a tweet can be used to detect spams, such as dash count in hostname, longest domain label, and URL count (Chen et al., 2014; Ma et al., 2009). However, links or URLs are typically shortened using services like TinyURL and bit.ly to stay within the character limit of tweets (Welch et al., 2011; Grier et al., 2010; Wang, 2010b). The shortened URL prevents making use of keywords or other interesting artifacts the original URL may contain, making additional processing of the URL necessary in order to extract features from the original URL.

In our study, we took the second approach by only looking at the content of the tweets and used machine learning techniques to detect emotions spams using human labeled data. This is because the user-based approach usually involves obtaining large number of historical information from a user. However, as discussed before, one practical reason for using emotion hashtags to retrieve data is its convenience, and we want to avoid bringing extra steps to this process. Thus we limit the scope of this study to the tweet content only.

3 Methodology

For our study, we collected tweets based on Plutchik's (1994) eight basic emotion types: *anger, anticipation, disgust, fear, happiness, sadness, surprise,* and *trust*. A total of 5,600 tweets were collected, 700 tweets per emotion type, and 200 of which were used for the pilot study. For each emotion type, we assigned 18 to 35 hashtags as seeds that have synonymous meaning to the emotion based on selected subset of Mohammad and Turney (2010) NRC Hashtag Emotion Lexicon (Appendix A.1).

3.1 Pilot Study

In our pilot study, we collected 200 tweets for each of the eight basic emotion types using emotion hashtags at a particular time point to briefly get a sense of how this strategy would work. We discovered that time has a major effect on the occurrence frequency of topics which can evoke a certain public emotion. For example, data collected in October 2016 were rich in topics about halloween, resulting in a higher frequency of fear emotion. This can cause bias towards the natural distribution of the emotions on Twitter.

3.2 Data Collection

To avoid tweets being concentrated on a specific topic, we decided to collect tweets from a wider time window for our final data. However, the Twitter official API has a strict constraint on time which only allows retrieving tweets that are within a week. Thus, we had to come up with an alternative way to collect tweets that are older. As a solution, we drew upon an open source Python project on GitHub - *GetOldTweets-python* (Henrique, 2016). The final Python code that we used was customized to add support for collection of emojis and URLs. With the code, a total of 4,000 tweets, 500 tweets per emotion type, between September 1st, 2015 to March 1st, 2016 were randomly collected for the eight basic emotion types. Each emotion type was crawled with selected subset of NRC Hashtag Emotion Lexicon (Mohammad and Turney, 2010).

3.3 Data Coding

With the all the tweets we gathered from the pilot study (200 tweets) and final data (500 tweets), each person manually annotated two emotion types. Thus, everyone in the group was responsible for annotating a total of 1400 tweets. We first identified whether the tweet expresses the emotion of its hashtag by coding either 1 or 0; 1 if it did express the hashtag emotion and 0 if it did not. Tweets that did not express its hashtag emotion were assigned 0s, and were further coded into one of the seven subcategories: negation, sarcasm, description, promotion, named entity, saying, and polysemy. These categories were developed by analyzing the tweets through a ground-up approach. Table 1 shows the description and examples we used as a rationale for annotating for each subcategory.

The seven subcategories were constructed from two dimensions: user intention and linguistic nature. The user intention subcategories include Promotion (*PR*), Saying (*SY*), and Description (*D*) as annotation of these subcategories depend on the user's—Tweeter's—intention. For these subcategories, the user's primary intention in tweeting the emotion spam was used in annotating the tweets. Thus, each tweet could only be assigned either none or one of these three subcategories. None of the user intention subcategories would be assigned to a tweet if the user's intention of an emotion spam could not be detected.

The subcategories that capture the linguistic nature of the tweets are Named Entity (*NE*), Polysemy (*PL*), Sarcasm (*SC*) and Negation (*NE*). Each emotion spam can be assigned multiple linguistic nature subcategories as long as a particular linguistic phenomenon was identified from the tweet. If an emotion spam does not have any linguistic phenomenon that can be captured by the four linguistic subcategories, no linguistic subcategories would be assigned to it. The seven subcategories that were identified from the data set were comprehensive enough to cover all the cases in our corpus. All emotion spams were assigned to at least one of the seven subcategories.

In addition to the text of the tweet, we also annotated the images that were attached to the tweets in order to analyze the correlation between them. The images were annotated on two parts. First part was annotating whether the image displayed the emotion of the hashtag or not by 1 and 0; 1 meaning yes it does display the emotion of the hashtag and 0 meaning no it does not. Second part is categorizing the image in one of the following classifications: Photo, Text, or Both. For example, if a tweet for the happiness emotion had a photo of a person smiling, then the image would be annotated with "1" and "Photo."

Remaining data fields of the tweet, such as the Tweeter's profile information were not gathered and analyzed due to the scope of this project. However, we did use the Tweeter's information, such as the username along with the original link of the tweet to confirm that our annotation are accurate. This was inevitable as we often faced difficulties in categorizing the tweets without the

Category	Code	Description	Example
Promotion	PR	Aims at advertising a product, service, event, beliefs, information, and personal social media account	• A few practical steps for dealing with #anger toward God $[URL]$ • #everyone #follow #me #today for #details On when I will be on #live $[URL]$ • $[URL]$ #Panic Attack #Solution That Works!
Saying	SY	Uses that expressions that generally contains advice or wisdom, such as quotes	• Life only comes around once so do whatever makes you #happy • Just because you're angry doesn't mean you have to react #Anger • #Everything you want is on the other side of #fear
Description	D	Uses emotion words to refer to a particular concept - describe something - than the emotion itself	• When Pumphrey gets #scary... • This is about having no friggin candidate and they are #Pissed • That's the #Sad part
Named Entity	NE	Uses emotion words that refer to a particular thing or concept other than the emotion itself	• Inside Out Themed party #fear #imsocute • Loved the #Joy #soundtrack • What will #Prospect say as UK Govt scraps CCS funding?
Polysemy	PL	Expresses coexistence of many possible meanings for a word or phrase	• That was a nice drink apparently #pissed #night • You need to see this guy in action! He got some #mad skills! • Well this is some very bizarre #tradeweek #content
Sarcasm	SC	Uses irony to mock or convey contempt	• More #racism from the realDonaldTrump campaign. What a #surprise • Entertainer gets paid for entertaining #shock! • watched the trailer for The Boss. I'm going to be saving a FORTUNE in cinema tickets next year. #sad
Negation	NG	Expresses emotion that contradicts the hashtag emotion	• #Rosy pulled the curtain into the cage to make a comfy bed. Its so cute I can't be #mad • #mad is so #good • Pretty sure it's impossible for me to be sad right now life is so gooood #life #sad

Table 1: Rationale for Each Sub-category

hashtag emotion since many were ambiguous to classify just by looking at the text. Some tweets were found out to be spams upon visiting the original link of the tweet or the link attached to the tweet. In order to perform consistent annotation in such high-level ambiguity, we cross-annotated each other's tweet annotations and reached 100% inter-annotator agreement on all the labeling of tweets.

4 Corpus Analysis

The main objective of this work is to identify different types of emotion spams and their characteristics. Moreover, the images posted with tweets were also analyzed in terms of their types and emotion. We address the following research questions in this section in order:

- What types of emotion spams are commonly found on Twitter? (*RQ1*)
- Which emotion hashtags are more likely to be used in emotion spams for each emotion type? (*RQ2*)
- How are emotion spams different from tweets that express a particular emotion in terms of tweet features (hashtags, URLs, emojis, keywords)? (*RQ3*)
- How does the emotion in the text associate with the image in the tweet? (*RQ4*)

4.1 Emotion Spams Distribution

Among the 4,000 tweets for all eight emotion types, 1,548 (38.70%) of them were marked as emotion spams. The high ratio of emotion spams contained in tweets that have emotion hashtags suggests the existence of excessive volume of noisy data in the data set labeled by emotion hashtags. The percentage of emotion spams is reasonable when compared with the result obtained from the work of Roberts et al. (2012). According to their study, 57% of the 7000 tweets they annotated were labeled as having no emotions, even though their data set was compiled with tweets that contained topic hashtags that they believed to frequently evoke emotions. The percentage of emotion spams in our data set is lower than their result because the tweets in our data set contain emotion-word hashtags.

The number of emotion spams vary across emotion types. As shown in Table 2, emotion type with the highest percentage of emotion spam is *anticipation* (243, 48.60%) and disgust has the fewest emotion spams (34, 6.80%).

Emotion Type	Sum	Percentage
anticipation	243	48.60%
trust	237	47.40%
sadness	231	46.20%
fear	229	45.80%
anger	215	43.00%
surprise	209	41.80%
happiness	150	30.00%
disgust	34	6.80%

Table 2: Number of Emotion Spams of Each Emotion Type

Seven subcategories of emotion spams were identified through a ground-up approach: promotion, saying, description, polysemy, named entity, sarcasm, and negation. Only a small number of emotion spams (79, 5.10% of emotion spams in total) were assigned more than one subcategory. Out of the 79 multi-subcategory emotion spams, 77 tweets were assigned two categories and only 2 were assigned three subcategories. The majority of the identified emotion spams belong to the promotion subcategory (1,155, 74.61% of emotion spams in total) and the distribution of emotion spams in all the subcategories is shown in Table 3.

Subcategory	Number	Percentage
Promotion	1,155	74.61%
Saying	145	9.37%
Description	133	8.59%
Polysemy	45	2.91%
Named Entity	110	7.11%
Sarcasm	23	1.49%
Negation	11	0.71%

Table 3: Number of Emotion Spams of Each Subcategory

To examine the distribution of emotion spams in different subcategories in depth, we analyzed the distribution of subcategories of emotion spams across the eight emotion types. A Chi-square test of association was performed and the result showed that the emotion types are strongly correlated with the distribution of emotion spam subcategory within an emotion type (χ^2 = 449.73, p $<$ .005). Figure 1 shows the major three subcategories of the emotion spams for each of the eight emotion types: promotion, saying, and description. It is important to note that these three major subcategories were mutually exclusive in the annotation process. Except for *anger*, the three subcategories together contribute to more than 90% of emotion spams for each emotion type. For *anger*, however, these three subcategories only account for 83.72% of the emotion spams. It can also be seen from Figure 1 that the emotion type *fear* has the most emotion spams of *saying* (57, 24.98% of emotion spams in *fear*). This is due to many tweets that contain hashtags that express fear as not fear itself, but something to overcome. For example, "*#Everything you want is on the other side of #fear*".

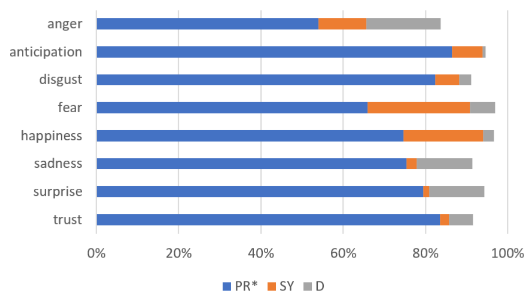

Figure 1: Distribution of Major Subcategories of Emotion Spams

PR = Promotion, SY = Saying, D = Description

The subordinate subcategories of emotion spams are polysemy, named entity, sarcasm, and negation. Figure 2 shows the distribution of these subordinate categories of emotion spam within each emotion type. *Trust* tweets have a large number of named entity emotion spams (44, 18.57% of emotion spams in *trust*). A further analysis into the *trust* tweets revealed that "#TrusTed" have been used for Ted Cruz's political campaign during the period when the tweets were crawled. The "#TrusTed" contributes to the majority of the named entity subcategory in the *trust* data. The use of "#Belief" to refer to a popular TV show also accounts for a large part of the named entity subcategory in the *trust* tweets. The emotion spams in *sadness* has many cases of polysemy (22, 9.52% of emotion spams in *sadness*) and sarcasm (18, 7.79% of emotion spams in *sadness*). This is mainly because "#sad" is often used to mean *pathetic* rather than to express the actual emotion of sadness. Also, in such cases, the authors were usually being sarcastic. The polysemy subcategory is comprised mainly of tweets that use "#mad" as being crazy rather than angry. For example, "*@batemanjason Just want you to know that you're an excellent actor. I have #mad #respect for your craft.*"

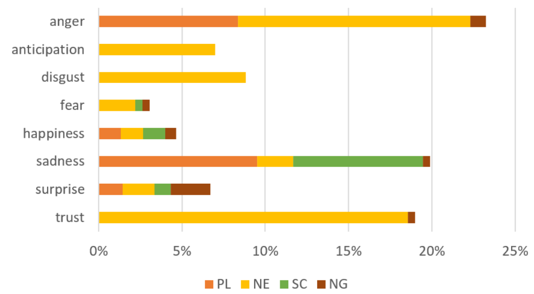

Figure 2: Distribution of Subordinate Sub-categories of Emotion Spams

PL = Polysemy, NE = Named Entity, SC = Sarcasm, NG = Negation

4.2 Critical Emotion-word Hashtags and Emotion-specific Keywords

4.2.1 Least Efficient Emotion-word Hashtags

The emotion-word hashtags we used to crawl tweets were a selected subset of the NRC Hashtag Emotion Lexicon generated by Mohammad and Turney (2010). Even though each word included in the NRC Hashtag Emotion Lexicon was given a score to suggest the word's degree of word-emotion association, the score was generated from noisy emotion labeled data obtained from *distant supervision*. The emotion-word hashtags used in this study were examined in terms of their percentage of introducing emotion spams, which is the number emotion spams divided by the number of tweets containing that particular emotion-word hashtag. The higher the percentage of emotion spam introduced by an emotion-word hashtag, the more careful researchers should be in using that emotion-word hashtag to get corresponding emotion-labeled tweets. Table 4 shows the emotion hashtags with relatively high percentage (> 20%) of introducing emotion spam for each emotion type. Only the emotion-word hashtags that have more than five occurrences in the corpus were considered in order to avoid the bias resulting from insufficient samples.

4.2.2 Emotion-specific Keywords

Mohammad (2012) and Mohammad and Kiritchenko (2015) revealed that the words used in

Emotion Type	Emotion-word hashtags and their likelihood of introducing emotion spams
anger	#anger(85.51%), #seeingred(75.00%), #mad(58.87%), #angry(45.28%)
anticipation	#await(86.96%), #prospect(72.81%), #expecting(64.65%), #expectation(60.47%)
disgust	N/A
fear	#fear(86.55%), #panic(62.86%), #nightmare(47.62%), #scared(34.92%), #creepy(24.49), #scary(21.74%)
happiness	#joy(50%), #smiling(43%), #happiness(37%), #delighted(25%), #happy(24%)
sadness	#depression(77.99%), #tragic(66.67%), #sadness(51.28%), #depressed(33.33%), #sad(25.12%)
surprise	#astounding(100.00%), #bombshell(88.10%), #surprising(57.89%), #wakeupcall(57.14%), #surprise(48.59%), #shock(31.25%)
trust	#trusted(64.86%), #belief(54.35%), #faith(48.84%), #trust(36.84%)

Table 4: Emotion-word Hashtags with High Probability of Introducing Spams

a tweet were found to be associated with the emotion expressed in the tweet. This is the basis for the hypothesis that the top keywords found in emotion spam and emotional tweets are different for each emotion type. Figure 3 lists the top five keywords appeared in the emotion spams and emotional tweets for the eight emotion types in descending order of their frequencies. Some categories have more than five words because the words share the same frequency. It is important to note that all emotion words used to crawl the data and all function words with little lexical meaning were removed from the list. Additionally, all words were normalized to their lower-case forms before the counting.

As Figure 3 shows, emojis are more prevalent in emotional tweets than emotion spams in general. This finding is consistent with the findings discussed in the previous section. The top five keywords for emotion spams and emotional tweets of the same emotion type have few overlaps. The only overlaps are "go" in *anticipation*, "love" in *happiness*, "how" in *sadness*, and "god" in *trust*. The word "just" is among the top five keywords of the emotional tweets in the six emotion types (*anger*, *disgust*, *fear*, *sadness*, *surprise*, *trust*), but is only listed among the top five words in *anger*. The difference of the top keywords suggests the possibility of the difference in linguistic styles exhibited in the tweets from different social and cultural groups.

Emotion Type	Emotion Spams	Emotional Tweets
anger	anger, got7, just, lonely, check, din, love	😡, get, just, can't, 😤
anticipation	baby, iceburghsociety, info, go, capital, pregnancy	see, new, excited, days, day, go
disgust	mean, coming, soonand, soon, tomorrow	like, people, just, really, one
fear	allah, halloween, get, horror, life, how, boo	halloween, just, know, people, dark
happiness	love, life, how, follow, new	love, 😍, ❤️, day, family, smile
sadness	anxiety, mentalhealth, how, help, kids, suicide	get, just, today, how, 😭
surprise	looks, beautiful, 💋, birthday, love, one	just, day, know, did, 😂, know
trust	god, broadcast, love, americans, claims	god, love, believe, life, hope, jesus, just

Figure 3: Top Keywords for Emotion Spam and Emotional Tweets

4.3 Tweet Features

4.3.1 Hashtags

The number of hashtags used per tweet is important to explore. It was found that the number of hashtags used by spammers was significantly higher than those by legitimate users (Antonakaki et al., 2016). Emotion spams should be studied separately when investigating its hashtags because it has a different concept from spam that is generally used. A t-test was performed upon the number of hashtags in a tweet against its presence of emotion. The result confirmed a statistically significant difference between the number of hashtags for emotion spams (M_0 = 4.41, SD = $\pm$2.91) and emotional tweets (M_1 = 2.65, SD = $\pm$2.16) [t'(2610.380) = 20.535, $p < .001$], suggesting that emotion spams have significantly more hashtags than emotion tweets.

The number of hashtags for each emotion type was further compared. As shown in Table 5, an average of 3.33 hashtags were included in each tweet among all the annotated tweets (4000 in total). Among the eight emotions, *happiness* has the most hashtags on average ($M_{happiness}$ = 4.7), whereas *disgust* has the fewest hashtags on average ($M_{disgust}$ = 2.15). The one-way ANOVA result showed that the emotion type had a significant effect on the number of hashtags [$F(7, 3992) = 44.983$, $p < .001$].

Emotion Type	# Hashtag	# URL	# Emoji
anger	2.73	0.32	0.37
anticipation	3.1	0.53	0.31
disgust	2.15	0.23	0.16
fear	3.63	0.52	0.22
happiness	4.7	0.57	0.65
sadness	3.29	0.41	0.28
surprise	3.21	0.45	0.4
trust	3.81	0.48	0.16
Average	3.33	0.44	0.32

Table 5: Number of Hashtags, URLs, and Emojis for Each Emotion Type

4.3.2 URLs

Tweets often contain different type of URLs for a number of reasons. A lot of previous studies on spam detection have already attempted to draw upon this feature to improve their detection accuracies (Chen et al., 2014; Yan et al., 2016; Antonakaki et al., 2016). However, these studies only made use of the presence of URLs and none have identified underlying URL types. During our data labeling process, we noticed that the content buried inside the URLs might play a role in the interpretation of a tweet. Thus, URL type can be an important feature to determine the nature of a tweet.

One of the barriers that hinder researchers from investigating deeper into the URL information for useful insights is the normalization step that is required. This is simply because the URLs are typically shortened to stay within the Twitter's character limit (Welch et al., 2011). A shortened URL needs to be unwrapped first before it can be further categorized.

An important contribution we made is that we took the first step to make use of the rich information buried in the URLs. We developed a URL classifier with Python which can automatically identify URLs in a given tweet and unwrap them as necessary using HTTP response header information. Then, we developed a set of rules to classify these URLs (see Table 6). It is important to note that if a URL is linking to a photo or a video on a social networking site, it is still considered as *Media* type rather tha *SNS* type.

We only drew upon the unwrapped URL to determine the content type because loading the web page and deciphering the HTML content is computationally much more expensive. Thus, it is not suitable for any large scale detection task. Besides, we were able to obtain very promising results simply by looking at features that comprise the URL.

The majority of the URLs are links to either media (36.91%) or articles (24.26%). Only 21.37% of the URLs were unidentified or did not belong to any of the four URL types, such as a link to an organization's homepage. Figure 4 further shows the distribution of the five URL types in each emotion. It is clear that *happiness* has more *Media* type URLs than other emotions and *anticipation* has the most unidentified URLs.

The number of URLs in a tweet was also examined in detail (Table 5). A t-test performed on the number of URLs against the presence of emotion revealed that there were significantly more URLs in an emotion spam (M_0 = 0.74, SD = $\pm$0.51) than in an emotional tweet (M_1 = 0.25, SD = $\pm$0.44) [t'(2930.790) = 31.064, $p < .001$]. A one-way ANOVA further suggested that there was a significant effect of emotion type on number of URLs [$F(7, 3992) = 25.628$, $p < .005$]. Among the eight emotion types, *disgust* ($M_{disgust}$ = 0.23) has much less URLs than other emotion types on average.

Type	%	Notes
Article	24.16%	links to any article
Media	36.92%	links to an image or a video
Shop	3.3%	links to online shops
SNS	14.25%	links to social media
Other	21.37%	anything that does not fit

Table 6: URL Types

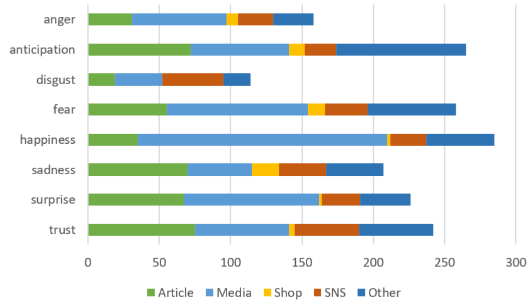

Figure 4: Distribution of URL Types in Each Emotion

Emotion	Photo	Text	P+T	Total
trust	24.77%	22.02%	53.21%	109
surprise	54.65%	2.33%	43.02%	86
sadness	49.43%	18.39%	32.18%	87
happiness	58.97%	10.26%	30.77%	117
fear	59.18%	10.20%	30.61%	98
disgust	80.00%	6.67%	13.33%	45
anticipation	49.38%	9.88%	40.74%	81
anger	63.49%	14.29%	22.22%	63
Total	52.48%	12.24%	35.28%	686

Table 7: Distribution of Image Types Per Emotion in Percentage (*P+T = photo + text, total is represented in count*)

4.3.3 Emojis

Previous studies suggest that emojis in tweets usually carry rich sentiment information (Novak et al., 2015) and some are closely related to specific semantic content (Barbieri et al., 2016). To further understand the associations between emojis and emotions, we conducted a *t*-test upon the number of emojis in the corpus against the presence of emotion. The result confirmed a statistically significant difference in the number of emojis for emotion spams (M_0 = 0.20, SD = $\pm$0.85) and emotional tweets (M_1 = 0.39, SD = $\pm$1.26) [t'(3981.447) = -5.732, $p < .005$], suggesting that tweets that do express the emotion indicated by its hashtags have significantly more emojis than those do not.

Moreover, the one-way ANOVA result showed that there was also a significant effect of emotion type on the number of emojis [$F(7, 3992)$ = 10.231, $p < .005$]. Among the eight emotion types, *happiness* has the highest mean of number of emojis ($M_{happiness}$ = 0.65), followed by *surprise* ($M_{surprise}$ = 0.40) and *anger* (M_{anger} = 0.37). This suggests that number of emojis in a tweet might be a promising feature for detecting emotion spam as well as identifying emotion type.

4.4 Emotion of Attached Images

One of our research questions is to compare the emotion expressed by the tweet text with the emotion expressed by the image attached to the text. When collecting the 4,000 tweets, we have also crawled the images attached to the tweets. The images were divided into three categories: 1) image that is purely a photo; 2) image that only contains text; 3) image that contains both photo and text. For tweets that have images, we further annotated whether the image contains the emotion assigned to the hashtag. Table 7 shows the number of images in each emotion type, and the distribution of the three categories. Overall, 17.15% of the tweets in the dataset have images. When broken down by emotion types, the result showed that some emotions contain more images than others. More than 20% of happiness and trust tweets have images whereas only 9% disgust tweets contain images. In terms of categories, over half of the images are purely photos, 35.4% contain both text and photos, and the remaining 12.2% contain only texts. Except for the *trust* emotion, this relative distribution holds consistent across most emotion types, with most of the images being photos and photos+text being the second most. However, for *trust*, the largest category is photo+text (50%) . *Trust* is an emotion that has high percentage of emotion spams (47.40%) as well as high percentage of promotional emotion spams (over 80%). A typical image found in the trust data set is represented by Figure 5.

Figure 5: Example of Image in Trust Data set

We observed a weak positive correlation on the emotion between tweet's texts and images (Spear-

C	Photo	Text	P+T	Total
True	47.50	36.90	35.95	42.13
False	52.50	63.10	64.05	57.87

Table 8: Consistency of Emotions Between Texts and Images in Percentage (C = consistency)

man's $r = .15$, $p < .001$). Table 8 shows that 57.97% of the tweets displayed consistent emotions between tweet texts and images. Namely, if the tweet text expresses the emotion assigned by the hashtag, the attached image will also express the same emotion, and vice versa. When looking at the results for each image category, we found that the consistency rate is higher for images that contain both text and photos (64.05%) and images that contain only text (63.10%). This might indicate that the presence of text in images results in higher emotion consistency rate.

5 Machine Learning Experiments

Because a large proportion of emotion spams in our corpus belong to the subcategory of promotion (74.61% of all emotion spams), we decided to focus on detecting emotion spams of this subcategory. If we are able to detect these promotional emotion spams, the total amount of emotion spams would be reduced dramatically. We first normalized the tweet content by conducting the following procedures: 1) Replace URL address with the string *url*. 2) Replace user mentions (represented by @username) with the string @*user*. 3) Remove all hashtags. 4) Change all characters to lower case. 5) Normalize words that have repeated character sequences and reduce the length to two (i.e. *waaaaaayyyyy* would be normalized to *waayy*). 6) Tokenize and stem the corpus. We used the Tweet-Tokenizer tokenizer that comes with the Natural Language Toolkit (NLTK), which is a free library for natural language processing. This tokenizer is able to handle elements, such as emoticons, emojis, URLs and HTML en-coding properly. 7) Remove stop words, which is also done by using the module provided by NLTK. For the experiments, we applied grid search function offered by Scikit Learn, a free machine learning library written in Python, to select optimal parameters automatically using a 10-fold cross-validation. Only 20% of the data was used for the testing of the model. Table 5 summarizes the results from the experiments.

In the first experiment, we replicated the method proposed in Benevenuto et al. (2010)'s work. Namely, we extracted the proposed features from tweets except for the number of times the tweet has been replied since this requires data from the users' account information. In total we have implemented 9 features and we also used support vector machine (SVM) classifier to perform the task. Scikit Learn was used to implement the machine learning algorithm. The result showed an accuracy (i.e.Micro F1 value) of 81.13%, which is much lower compared to the result reported in the original study (i.e. 87.2%). Moreover, the model was only able to detect 67.42% of the promotional emotion spams (78.5% was reported in the original study). This indicates that the classifier has difficulty in identifying promotional emotion spam identified in this study. As discussed before, the promotional content existed in emotion tweets are not exactly the same as the spam in the traditional sense and the number of times the tweet has been replied was not used in this experiment. It is not very surprising to us that the model did not perform as well as when tested on our data set.

Features	A	F	TP	FP
9-F	81.13	81.35	67.42	12.01
14-F	84.50	83.53	**86.14**	15.93
Unigram	86.88	86.59	82.94	11.71
14F∪Uni	**88.13**	**87.87**	85.31	**10.86**

Table 9: Features, Accuracy (*A*), and F-measure, True Positive rate (*TP*), False Positive rate (*FP*), for promotional spam classifiers using different feature sets. For each column in the table the best values (highest *TP*, *A* and F-measure, and lowest *FP*) are shown in bold.

In aforementioned analysis of Tweet features, we discussed the importance of the semantic meanings carried by URLs, and developed a method to categorize them. We also found that there is a significant difference between the number of emojis in emotional tweets and emotion spams. Based on the result of our analysis we added the following features to the model besides the existing 9 features: 1) the URL types 2) number of emojis 3) number of emojis per word 4) whether the tweet starts or ends with a hashtag 5) whether the tweet starts or ends with a URL. In total, we have implemented 14 features. The new model was put into test and SVM was again used as we think using same classifier would help us to interpret and compare the results obtained from the different models. The result showed that the 14-Feature model has an overall higher accuracy compared with the 9-Feature model (84.50%). Moreover, the 14-Feature model has improved significantly in terms of correctly identifying promotional emotion spams—The true positive rate has increased from 67.42% to 86.14%. However, the false positive rate (i.e.wrongly classified non-promotional tweet as promotional spams) also increased from 12.01% to 15.93%. In fact, when compared with results yielded from all other experiments, the 14-Feature model scores highest in terms of correctly identifying promotional spams. This indicates that this model is good at detecting promotional content.

In the third experiment, we implemented a basic unigram BOW model on the normalized corpus. We extracted all unique terms from the corpus and added them as features. When calculating the occurrence of a term, only its binary occurrence (i.e. 0 or 1) was considered. The unigram model showed a result of 86.88% accuracy, which is higher than the accuracy obtained from the 14-Feature model. This model also performed well in terms of false positive rate and the rate of 11.71% is lower than both the 9-Feature and 14-Feature model. However, the true positive rate is lower than that obtained from the 14-Feature model (82.94% compared with 86.14%).

Finally, in the last experiment we combined the 14-Feature model and the unigram model. Since the 14-Feature model scored high in true positive rate while the unigram model performed well in terms of false positive rate, these two models seemed to be complementary to each other. The final model implemented a linear SVM with C = 0.1. The result showed an accuracy of 88.13% and the overall F-measurement is 87.87%. Both yielded the best result when compared with the other three models. The average accuracy obtained during the 10-fold cross-validation is 87.53% with a standard deviation of 1.41%, which suggests that the model is also robust. The combined model also performed best in terms of false positive rate and 10.86% was the lowest among the four models. However, although very close, the true positive rate obtained by this model is still slightly lower than that obtained by the 14-feature model—85.31% compared with 86.14%

SVM is one of the most commonly used classifiers both in the domain of emotion analysis and spam detection (Benevenuto et al., 2010; Martinez-Romo and Araujo, 2013; Mohammad and Kiritchenko, 2015; Yan et al., 2016). Other commonly used classifiers include Naive Bayes (Freeman, 2013; Wang, 2010b), K-Nearest Neighbor (Hasan et al., ; Miller et al., 2014), and Random Forest (Mccord and Chuah, 2011). Mccord and Chuah (2011)'s study implemented all these classifiers in a spam detection task and the Random Forest classifier yielded best result. In our experiments, we only used the SVM classifier. It would be interesting to implement the same features with different classifiers and compare the results. The model developed by us has been made freely accessible online (Yang et al., 2016) and is ready to be used by researchers who may be interested.

6 Conclusion

In this study, we aimed to understand the characteristics of emotion spams and to develop a machine learning model to identify them from any emotion-word containing Twitter data. Through human-labeling, we have constructed a data set of 4,000 tweets that contains both emotional tweets and emotion spams. By analyzing this data set, we observed a moderate proportion of emotion spams (38.70%). We discovered that promotion, sayings, and description are the most common causes for a tweet to be assigned as an emotion spam. Among the three categories, promotion takes up the majority of emotion spams across all emotion types. Other possible causes include named entities, polysemy, sarcasm, and negation. The emotion types are found to be strongly correlated with the distribution of emotion spam subcategory within an emotion type. We further presented a list of least efficient emotion-word hashtags that are often used to obtain Twitter data and a list of critical keywords for both emotional tweets and emotion spams for each emotion type.

We also examined various features of a tweet and discovered that the number of hashtags, emojis, URLs, and URL types are all very important when it comes to detecting emotion spams. This finding was further confirmed by the higher detection accuracies after adding them as features into our machine learning model. Images attached to the tweets as well as the consistency of emotions between the tweet's text and images have been studied in detail as well. We found a weak positive correlation between the emotion of the tweet's texts and images and also observed that the presence of text in images tends to contribute to higher emotion consistency. Finally, we developed a machine learning model that was able to reach an accuracy of 88.13%, and the model has been made freely accessible by others.

Our study has several limitations. Studies have found that multiple emotions can be expressed by a tweet (Roberts et al., 2012; Yan and Turtle, 2016). However, we could analyze only the single emotion that is expressed by the emotion-word hashtag that was used to crawl the data. Secondly, additional tweets for each emotion type would have been annotated if more time and resources were available. Third, labeling emotion by trained researchers other than the Twitters who tweeted the tweets can bear a certain level of subjectivity and inaccuracy even though all the tweets were labeled by two researchers and reached 100% inter-annotator agreement. Finally, the features of Twitters were not taken into account in our study. It is one of our future works to analyze the information of the Twitters of emotion spams and utilize the features of Twitters in detecting emotion spams.

References

Despoina Antonakaki, Iasonas Polakis, Elias Athanasopoulos, Sotiris Ioannidis, and Paraskevi Fragopoulou. 2016. Exploiting abused trending topics to identify spam campaigns in twitter. *Social Network Analysis and Mining*, 6(1):1–11.

Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. 2016. What does this emoji mean? a vector space skip-gram model for twitter emojis. In *Language Resources and Evaluation conference, LREC, Portoroz, Slovenia*.

Natalya N Bazarova, Yoon Hyung Choi, Victoria Schwanda Sosik, Dan Cosley, and Janis Whitlock. 2015. Social sharing of emotions on facebook: Channel differences, satisfaction, and replies. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 154–164. ACM.

Fabricio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida. 2010. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, volume 6, page 12.

Leyla Bilge, Thorsten Strufe, Davide Balzarotti, and Engin Kirda. 2009. All your contacts are belong to us: automated identity theft attacks on social networks. In *Proceedings of the 18th international conference on World wide web*, pages 551–560. ACM.

Chia-Mei Chen, DJ Guan, and Qun-Kai Su. 2014. Feature set identification for detecting suspicious urls using bayesian classification in social networks. *Information Sciences*, 289:133–147.

Munmun De Choudhury, Michael Gamon, and Scott Counts. 2012. Happy, nervous or surprised? classification of human affective states in social media. In *ICWSM*.

David Mandell Freeman. 2013. Using naive bayes to detect spammy names in social networks. In *Proceedings of the 2013 ACM workshop on Artificial intelligence and security*, pages 3–12. ACM.

Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. 2010. @ spam: the underground on 140 characters or less. In *Proceedings of the 17th ACM conference on Computer and communications security*, pages 27–37. ACM.

Maryam Hasan, Elke Rundensteiner, and Emmanuel Agu. Emotex: Detecting emotions in twitter messages.

Jefferson Henrique. 2016. Getoldtweets-python. `https://github.com/Jefferson-Henrique/GetOldTweets-python`.

Jeff Huang, Katherine M Thornton, and Efthimis N Efthimiadis. 2010. Conversational tagging in twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, pages 173–178. ACM.

FA Kunneman, CC Liebrecht, and APJ van den Bosch. 2014. The (un) predictability of emotional hashtags in twitter. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 26–34. [sl]: Association for Computational Linguistics.

Justin Ma, Lawrence K Saul, Stefan Savage, and Geoffrey M Voelker. 2009. Beyond blacklists: learning to detect malicious web sites from suspicious urls. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1245–1254. ACM.

Juan Martinez-Romo and Lourdes Araujo. 2013. Detecting malicious tweets in trending topics using a statistical analysis of language. *Expert Systems with Applications*, 40(8):2992–3000.

Michael Mccord and M Chuah. 2011. Spam detection on twitter using traditional classifiers. In *International Conference on Autonomic and Trusted Computing*, pages 175–186. Springer.

Zachary Miller, Brian Dickinson, William Deitrick, Wei Hu, and Alex Hai Wang. 2014. Twitter spammer detection using data stream clustering. *Information Sciences*, 260:64–73.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

Saif M Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

Saif M Mohammad and Peter D Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics.

Saif M Mohammad. 2012. # emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 246–255. Association for Computational Linguistics.

Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. 2010. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 189–192. ACM.

Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of emojis. *PloS one*, 10(12):e0144296.

Matthew Purver and Stuart Battersby. 2012. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 482–491. Association for Computational Linguistics.

Ashequl Qadir and Ellen Riloff. 2014. Learning emotion indicators from tweets: Hashtags, hashtag patterns, and phrases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics*, pages 1203–1209.

Kirk Roberts, Michael A Roach, Joseph Johnson, Josh Guthrie, and Sanda M Harabagiu. 2012. Empatweet: Annotating and detecting emotions on twitter. In *LREC*, pages 3806–3813. Citeseer.

Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. 2010. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 841–842. ACM.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2012. Harnessing twitter" big data" for automatic emotion identification. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 587–592. IEEE.

Alex Hai Wang. 2010a. Detecting spam bots in online social networking sites: a machine learning approach. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pages 335–342. Springer.

Alex Hai Wang. 2010b. Don't follow me: Spam detection in twitter. In *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*, pages 1–10. IEEE.

Michael J Welch, Uri Schonfeld, Dan He, and Junghoo Cho. 2011. Topical semantics of twitter links. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 327–336. ACM.

Jasy Liew Suet Yan and Howard R Turtle. 2016. Exploring fine-grained emotion detection in tweets. In *Proceedings of NAACL-HLT*, pages 73–80.

Jasy Liew Suet Yan, Howard R. Turtle, and Elizabeth D. Liddy. 2016. Emotweet-28: A fine-grained emotion corpus for sentiment analysis. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Qian Yang, Ziyao Chen, Yu Zhao, and Alex Hur. 2016. Tweet-emotion-spam-detection. `https://github.com/IAMQIANYANG/tweet-emotion-spam-detection`.

A Appendix

A.1 List of Synonyms for Each Emotion Types

Emotion	Synonyms
anger	angry, angered, angrily, angriness, anger, angrier, angriest, displeasure, crossness, irascibility, illtempered, slowburn, irate, mad, irked, infuriated, inatemper, choleric, upinarms, inhighdudgeon, foamingatthemouth, doingaslowburn, inalather, fittobetied, seeingred, bentoutofshape, tickedoff, teedoff, pissed, pissedoff, badtempered, acrimonious
anticipation	anticipated, anticipating, anticipation, anticipatable, expecting, anticipator, expectation, expectance, expectancy, prospect, lookingforwardto, lookforwardto, await, awaiting, countingthedays, lickingmylips, can'twait, cantwait, cannotwait
disgust	disgusting, disgustedly, disgusted, sickening, nauseating, nauseatic, repulsive, turnmystomach, gross, pukeable, discusting, discusted, discust, disgust
fear	fearsome, fearfully, fearfulness, fearful, fear, panic, agitation, dismay, givemethecreeps, givesmethecreeps, phobia, bugbear, nightmare, neurosis, scared, scaredstiff, scaredtodeath, petrified, alarmed, panicky, trembling, quaking, cowed, daunted, timid, timorous, fainthearted, twitchy, trepidatious, inacoldsweat, abundleofnerves, spooked, creepy, scary
happiness	happy, happiness, happily, happier, happiest, feelinghappy, joyfully, joyfulness, joyously, joyousness , joyous, joyful, joy, cheerful, cheery, jovial, jolly, jocular, gleeful, carefree, delighted, smiling, grinning, ingoodspirits, inagoodmood, lighthearted, pleased, satisfied, gratified, buoyant, feelingsunny, blithe, beatific, exhilarated, blissful
sadness	sadness, sadden, sadly, sadder, saddest, saddeningly, sad, unhappiness, depression, despondency,wretchedness, gloom, gloominess, unhappy, dejected, depressed, downcast, feelingdown, despondent, disconsolate, wretched, glum, gloomy, dismal, forlorn, crestfallen, inconsolable, feelingblue, downinthemouth downatthemouth, downinthedumps, tragic, heartbreaking
surprise	surprising, surpriser, surprisedly, surprised, surprise, astonished, astonishedly, astonisher, astonishment, astonishingly, astonishing, astonish, boltfromtheblue, bombshell, eyeopener, wakeupcall,shocker, startled, shocked, shock, takenaback, stupefied, dumbfounded, dumbstruck, bowledover, flabbergasted, astounding, staggering, eyeopening
trust	trustable, trusty, trustability, trustful, trusting, trusties, trustier, trustiest, trusted, trustily, trustiness, trust, belief, faith, certainty, assurance, credence, reliance