EMOTION ANALYSIS OF TWITTER DATA THAT USE EMOTICONS AND EMOJI IDEOGRAMS

Wiesław Wolny *wieslaw.wolny@uekat.pl*
University of Economics in Katowice
Katowice,Poland

Abstract

Twitter is an online social networking service on which users worldwide publish their opinions on a variety of topics, discuss current issues, complain, and express many kinds of emotions. Therefore, Twitter is a rich source of data for opinion mining, sentiment and emotion analysis. This paper focuses on this issue by analysing symbols called emotion tokens, including emotion symbols (e.g. emoticons and emoji ideograms). According to observations, emotion tokens are commonly used in many tweets. They directly express one's emotions regardless of his/her language, hence they have become a useful signal for sentiment analysis in multilingual tweets. The paper describes the approach to extending existing binary sentiment classification approaches using a multi-way emotions classification.
Keywords: Twitter, data mining, sentiment analysis, emotion analysis.

1. Introduction

Microblogging websites such as Twitter (`www.twitter.com`) have evolved to become a great source of various kinds of information. This is due to the nature of microblogs on which people post real-time messages regarding their opinions on a variety of topics, discuss current issues, complain, and express many kinds of emotions. As the audience of microblogging platforms and social networks grows every day, data from these sources can be used in opinion mining, sentiment and emotion analysis tasks. Opinions and related concepts such as sentiments and emotions are the subjects of study of sentiment analysis and opinion mining. The inception and rapid growth of the field coincide with those of the social media on the Web, e.g., reviews, forum discussions, blogs, microblogs, Twitter, and social networks. Most Natural Language Processing (NLP) methods perform without particular success in the social media. Almost all forms of social media are very noisy and full of all kinds of spelling, grammatical, and punctuation errors.

Current sentiment analysis methods typically focus on the polarity of like/dislike emotions. Sometimes neutral emotions can be detected in between. Human emotions are far beyond these simple metrics and are much more diverse. This implies that such polarity analysis gives only limited information on the actual intent of the author of the message. Defining either positive or negative emotions only is relatively simple, yet defining a complete and clear set of emotions is much more difficult. Researchers have thus created a wide range of research tools on the identification of basic emotions.

2. Related Research

Sentiment analysis is a growing area of the Natural Language Processing task at many levels of granularity. Starting from being a document level classification task ([33], [24]), it has been handled at the sentence level ([16], [18]) and more recently at the phrase level ([34], [6]), or even at the polarity of words and phrases ([15], [12]).

However, the informal and specialised language that is used in tweets as well as the nature of the microblogging domain make sentiment analysis in Twitter a very different task. With the growing number of blogs and social networks, opinion mining and sentiment analysis have become fields of interest to many researches. A very broad overview of the existing work was presented in [25]. J. Read, in [27], used emoticons such as ":-)" and ":- (" to form a training set for sentiment classification. For this purpose, the authors collected texts containing emoticons from Usenet newsgroups. The dataset was divided into "positive" (texts with happy emoticons) and "negative" (texts with sad or angry emoticons) samples.

Researchers have also begun to investigate various ways of automatically collecting training data. Several researchers have relied on emoticons to define training data ([23], [9]). Barbosa and Feng in [7] exploited existing Twitter sentiment sites to collect training data. Davidov, Tsur and Rappoport [11] also used hashtags to create training data but they limited their experiments to sentiment/non-sentiment classification rather than the multi-way emotion classification that is presented in this article.

Extending research to many different kinds of emotions is a very new concept and has not been extensively studied yet. There are currently few examples in which researchers have gone beyond the polarity of sentiment analysis. Socher et al. [30] predicted five different dimensions of sentiments. Tromp and Pechenizkiy [32] used Pluchnik's wheel of emotions model and a rule-based approach for emotion detection in the text. A slightly different but also interesting approach was presented by Mohammad [20], who detected emotion on twitter posts by using emotion-word hashtags.

3. From Sentiment to Emotion Analysis

Sentiment analysis, which is also known as opinion mining, focuses on discovering patterns in the text that can be analysed to classify sentiment in that text. The term sentiment analysis probably first appeared in [21], and the term opinion mining first appeared in [10]. However, research on sentiments and opinions appeared earlier.

According to Liu "sentiment analysis is the field of study that analyses people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, and their attributes. It represents a large problem space. There are also many names and slightly different tasks, e.g., sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc." [19]. Sentiment analysis has grown to be one of the most active research fields in natural language processing. It is also widely studied in data mining, Web mining and text mining. In fact, it has spread from computer science to management sciences and social sciences due to its importance to business and society.

Sentiment analysis is predominantly implemented in software which can autonomously extract emotions and opinions from a text. It has many real-world applications, e.g. it allows companies to analyse how their products or brand is being perceived by their consumers, and politicians may be interested in knowing how people plan to vote in elections, etc. It is difficult to classify sentiment analysis as one specific field of study as it incorporates many different areas, such as linguistics, Natural Language Processing, and Machine Learning or Artificial Intelligence. Since the majority of sentiment that is uploaded to the Internet is of an unstructured nature, it is a difficult task for computers to process it and to extract meaningful information from it. Some of the most effective machine learning algorithms, e.g. support vector machines, naïve Bayes and conditional random fields, often produce no human understandable results.

Emotions are closely related to sentiments. Emotions can be defined as subjective feelings and thoughts. People's emotions have been categorised into distinct categories. Emotion analysis can thus be done as an additional layer on top of the (relatively) simpler sentiment classification. However, there is still no set of agreed basic emotions among researchers. Based on

[26], people have six primary emotions, i.e., love, joy, surprise, anger, sadness, and fear, which can be sub-divided into many secondary and tertiary emotions. Each emotion can also have different intensities.

Emotions in virtual communication differ in a variety of ways from those in face-to-face interactions due to the characteristics of computer-mediated communication, which may lack many of the auditory and visual cues that are normally associated with the emotional aspects of interactions. While text-based communication eliminates audio and visual cues, there are other methods for adding emotion. Emoticons, or emotional icons, can be used to display various types of emotions. Ortony and Turner [22] collated a wide range of research on the identification of basic emotions (Table 1).

Table 1. Identification of basic emotions

Theorist	**Basic Emotions**
Plutchik	acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise
Arnold	anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness
Ekman, Friesen and Ellsworth	anger, disgust, fear, joy, sadness, surprise
Frijda	desire, happiness, interest, surprise, wonder, sorrow
Gray	rage and terror, anxiety, joy
Izard	anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise
James	fear, grief, love, rage
McDougall	anger, disgust, elation, fear, subjection, tender-emotion, wonder
Mowrer	pain, pleasure
Oatley and Johnson-Laird	anger, disgust, anxiety, happiness, sadness
Panksepp	expectancy, fear, rage, panic
Parrott	love, joy, surprise, anger, sadness, fear
Tomkins	anger, interest, contempt, disgust, distress, fear, joy, shame, surprise
Watson	fear, love, rage
Weiner and Graham	happiness, sadness

There are many common categories in the different research studies as presented in Table 1, but there are still too many very different emotions for effective analysis. Some concepts aim to minimise the number of basic emotions. Jack, Garrod and Schyns [17] analysed the 42 facial muscles that shape emotions on the face and they came up with only four basic emotions, yet there is still no consensus on the basic set of emotions that would be generally accepted and could be objectively verified. For the purposes of this work, emotions can be classified into emoticon types similar to those in Wikipedia [5] (Table 2):

4. Twitter Data Description

Twitter has its own conventions that renders it distinct from other textual data; Twitter messages are called tweets. Twitter also has its own conventions that renders it distinct from other textual data. There are some particular features that can be used to compose a tweet 1.

The first pieces of information, UE Katowice and @UE_Katowice are the twitter name for the University of Economics in Katowice; #UEKatowice, #international, and #Erasmus are tags provided by the user for this message, the so-called hashtags. Users of Twitter use the "@"

Table 2. Sentiment and Emotion expressed by emoticons and emoji

Sentiment Polarity	Emotion Classes	Emoticons and Emoji ideograms related
Positive	Happiness Laugh Grin	:-) :) :D :o) :] :3 :c) :> =] 8) 😊😀
Negative	Sadness	>:[:-(:(:-c :c :-< :< :-[:[:{ 😢😩😞😥😿
	Anger	:-|| :@ >:(😠😤😡💢
	Cry	:'-(:'(😭😰😢😓😪😓
	Horror	😱
	Disgust Great dismay	D:< D: D8 D; D= DX v.v D-' 😝😷😒💩🍙
	Skeptical	😒
	Annoyed	😒😤😑😪
	Undecided	>: >:/ :-/ :-. :- : =/ = :L =L :S >.<
	Uneasy Hesitant	
	Cool	|;-) 😎👊
	Bored/yawning	🐻🚶😶😪
Neutral	Straight face No expression Indecision	:| :-|
	Suprise	>:O :-O :O :-o :o 8O O_O o-o O_o o_O o_o O-O 😱😳
	Shock	😯😦😧😨😵🤯😳

symbol to refer to other users. Referring to other users in this manner automatically alerts them. Users usually use hashtags to mark topics. This is primarily done to increase the visibility of their tweets. These symbols provide an easy way of identifying Twitter user names and topics and thus allows to search for and filter information on any subject. In the tweet the following emoticons, – :), and emoji characters — 😍🌍🛫💉 were used.

Twitter messages have many unique attributes, which differentiates twitter analysis from other fields of research. The first attribute is length. The maximum length of a Twitter message is 140 characters. The average length of a tweet is 14 words [14]. This is very different from the domains of other research studies, which were mostly focused on reviews which consisted of multiple sentences. The second attribute is the availability of data. With the Twitter API or other tools it is much easier to collect millions of tweets for training.

4.1. Emoticons

There are two fundamental data mining tasks that can be considered in conjunction with Twitter data, i.e. text analysis and symbol analysis. Due to the nature of this microblogging service (quick and short messages), people use acronyms, make spelling mistakes, and use emoticons and other characters that express special meanings. Emoticons constitute a metacommunicative pictorial representation of a facial expression that is pictorially represented by using punctuation and letters or pictures; they express the user's mood.

The use of emoticons can be traced back to the 19th century. The first documented person to have used the emoticons :-) and :-(on the Internet was Scott Fahlman from Carnegie Mellon University in a message dated 19 September 1982 [1].

Some emoticons as characters are included in the Unicode standard, three in the Miscella-

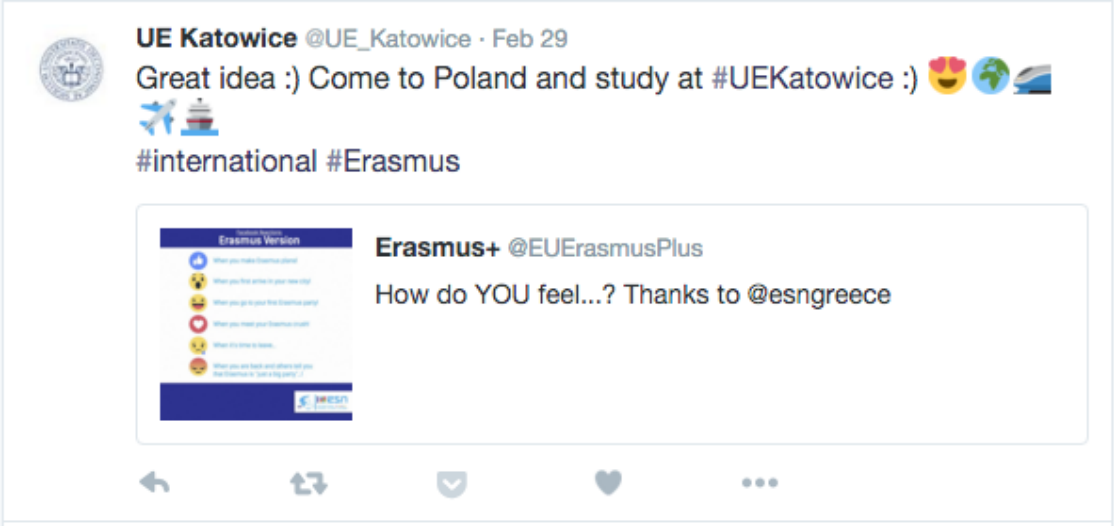

Figure 1. Example of a Tweet

neous Symbols block, and over sixty in the Emoticons block [4]. More symbols and meanings which can be used to determine emotional state can be found on the Wikipedia website [5]. The top 20 emoticons collected from 96 269 892 tweets is presented in [8].

4.2. Emoji ideograms

Emoji were originally used in Japanese electronic messages and spread outside of Japan. The characters are used much like emoticons, although a wider range is provided. The rise in the popularity of emoji is due to its being incorporated into sets of characters available in mobile phones. Apple in IOS, Android and other mobile operating systems included some emoji character sets. Emoji characters are also included in the Unicode standard [4]. Emoji can be categorised into similar categories as emoticons. Emoji can even be translated into English by using `http://emojitranslate.com` website.

5. Architecture of Emotion Classification System

The main problem is how to extract the rich information that is available on Twitter and how to use it to draw meaningful insight. To achieve this, at first stage of works an accurate analyser for tweets was build. Twitter allows developers to collect data via Twitter REST API [2] and The Streaming API [3]. Twitter has numerous regulations and rate limits imposed on its API, and for this reason it requires that all users must register an account and provide authentication details when they query the API. One of the best ways of connecting to Twitter Streaming API and downloading the data is by using Python and a library called Tweepy [28].

Twitter API exports data only in JSON format, which should be translated into readable for databases or an analytical software format. A combination of Twitter API, scripts for converting JSON to CSV [29], SAS Macro [13] or Excel Macro [31] was used to extract information from Twitter and to create an input dataset for the analysis. The entire process of data acquisition can be fully automated by scheduling the run of VBA or SAS macros.

Data can also be stored directly in JSON format in an NoSQL Database, which provides a mechanism for the storage and retrieval of data which is modelled in means other than the tabular relations used in relational databases. MongoDB (`https://www.mongodb.org/`) allows to store data in JSON-like documents with dynamic schemas (MongoDB calls the format

BSON), thus making the integration of data in certain types of applications easier and faster.

Since opinions have targets, further preprocessing and filtering of collected data was done using @twitter_names and #hashtags as targets in the way described in [2]. This method is more precise and provides better results than other text mining approaches. Software used for data analysis can be SAS Text Miner, SAS Visual Analytics or other tools. SAS Visual Analytics allows for direct import of Twitter data, but in order to use SAS Text Miner and other tools, the data have to be downloaded and converted.

Instead of using commercial software, all sentiment and emotion analysis tasks was solved using python programming language. Python language has many machine learning and data mining extensions that are suitable for this kind of work. The challenge remains to fetch customised Tweets and to clean data before any text or symbol mining takes place.

A classification of tweets was made in unsupervised learning method by using the lexicon-based approach. The sentiment lexicon contains a list of emoticons and emoji ideograms based on Table 2. Data was gathered by searching Twitter posts using Twitter API. An assumption must be made in order to use this method, this assumption is that the emoticon in the tweet represents the overall emotion contained in that tweet. This assumption is quite reasonable as the maximum length of a tweet is 140 characters, so in the majority of cases the emoticon will correctly represent the overall sentiment of that tweet. This kind of evaluation is commonly known as the document-level sentiment classification because it considers the whole document as a basic information unit. The model can be developed on a sample of data; then can be used to classify the emotions of the tweet.

First verification of the method used here indicates that the most recognisable are only the basic emotions. The obtained results are similar to work [8], in which Nick Berry reveals that the top 20 emoticons accounted for 90% of all 96,269,892 Tweets. Therefore, for effective emotion analysis we could use only a limited subset of emoticons and recognised only basic, common emotions, such as:

- happy – :) :D :-) =) (: =D D: :] 😊😀
- sad, unhappy – :(:-(-(😢😩😞😥😿
- undecided, sceptical – :/ =/ 😒
- surprise, shock – :o 😱😳

This set of emoticons and the corresponding emotions covers nearly 90% of all occurrences.

However objective verification of emotions is very difficult. Further work should go in the direction of carrying out supervised check what part of emoticons does not reflect correctly emotions of tweet.

6. Conclusions

Microblogging such as on Twitter has today become one of the major types of communication. The large amount of information contained in these websites makes them an attractive source of data for opinion mining and sentiment analysis. Most text-based methods of analysis may not always be useful for sentiment analysis in these domains. We as researchers still need novel ideas to make significant progress in this area. Using symbol analysis that makes use of emoticons and emoji characters can significantly increase precision in recognising many kinds of emotions. Also, applying Twitter names and hashtags to filter collected training data can provide better results. The most successful algorithms will probably be integration of natural language processing methods and symbol analysis

References

1. :-) turns 25. https://web.archive.org/web/20071012051803/http://www.cnn.com/2007/TECH/09/18/emoticon.anniversary.ap/index.html, 09 2007.
2. Twitter rest api. the search api. https://dev.twitter.com/rest/public/search, 05 2015.
3. Twitter the streaming apis. https://dev.twitter.com/streaming/overview, 05 2015.
4. Unicode miscellaneous symbols. http://www.unicode.org/charts/PDF/U2600.pdf, 05 2015.
5. List of emoticons. https://en.wikipedia.org/wiki/List_of_emoticons, 04 2016.
6. Apoorv Agarwal, Fadi Biadsy, and Kathleen R. Mckeown. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 24–32, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.
7. Luciano Barbosa and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 36–44, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
8. N. Berry. Datagenetics. http://www.datagenetics.com/blog/october52012/index.html, 05 2015.
9. Albert Bifet and Eibe Frank. Sentiment knowledge discovery in twitter streaming data. In *Proceedings of the 13th International Conference on Discovery Science*, DS'10, pages 1–15, Berlin, Heidelberg, 2010. Springer-Verlag.
10. Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pages 519–528, New York, NY, USA, 2003. ACM.
11. Dmitry Davidov, Oren Tsur, and Ari Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 241–249, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
12. Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06*, pages 417–422, 2006.
13. S. Garla and G. Chakraborty. tweets. Paper 324-2011. Oklahoma State University, Stillwater, OK, 2001.
14. Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *Processing*, pages 1–6, 2009.
15. Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98, pages 174–181, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics.
16. Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA, 2004. ACM.
17. Rachael E. Jack, Oliver G.B. Garrod, and Philippe G. Schyns. Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current Biology*,

24(2):187–192, 2014.
18. Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.
19. Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012.
20. Saif Mohammad. #emotional tweets. In **SEM 2012: 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 (SemEval 2012)*, pages 246–255, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics.
21. Tetsuya Nasukawa and Jeonghee Yi. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2Nd International Conference on Knowledge Capture*, K-CAP '03, pages 70–77, New York, NY, USA, 2003. ACM.
22. Andrew Ortony and Terence J Turner. What's basic about basic emotions? *Psychological review*, 97(3):315, 1990.
23. Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).
24. Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.
25. Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January 2008.
26. W Parrott. *Emotions in social psychology: Essential readings.* Psychology Press, 2001.
27. Jonathon Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, ACLstudent '05, pages 43–48, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
28. Matthew A. Russell. *Mining the Social Web, 2nd Edition Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. O'Reilly Media, 2013.
29. Falko Shultz. How to import twitter tweets in sas data step using oauth 2 authentication style. `http://blogs.sas.com/content/sascom/2013/12/12/how-to-import-twitter-tweets-in-sas-data-step-using-oauth/-2-authentication-style/`, 12 2013.
30. Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 151–161, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
31. Analytics Tools. Can sas be used to gather sentiments from twitter? `http://www.analytics-tools.com/2012/06/social-media-analytics-twitter-text.html`, 07 2012.
32. Erik Tromp and Mykola Pechenizkiy. Pattern-based emotion classification on social media. In *Advances in Social Media Analysis*, volume 602, pages 1–20. Springer, 2015.
33. Peter D. Turney. Thumbs up or thumbs down?: Semantic orientation applied to un-

supervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

34. Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.