Hierarchical Classification Approach to Emotion Recognition in Twitter

Conference Paper · December 2012

DOI: 10.1109/ICMLA.2012.195

3 authors:

Ahmed Esmin
Universidade Federal de Lavras (UFLA)

Roberto Lourenço Oliveira Júnior
Microsoft, Vancouver, Canada

Stan Matwin
University of Ottawa

Hierarchical Classification Approach to Emotion Recognition in Twitter

Ahmed A. A. Esmin
Department of Computer Science
Federal University of Lavras,
MG, Brazil
School of Electrical Engineering
and Computer Science
University of Ottawa, Ottawa, ON,
Canada
e-mail: ahmed@dcc.ufla.br

Roberto L. de Oliveira Jr.
Department of Computer Science
Federal University of Minas
Gerais, MG, Brazil
e-mail: robertolojr@dcc.ufmg.br

Stan Matwin
School of Electrical Engineering
and Computer Science,
University of Ottawa, Ottawa, ON,
Canada
Institute of Computer Science
Polish Academy of Sciences,
Warsaw, Poland
e-mail:stan@eecs.uottawa.ca

Abstract—**Twitter is a microblogging service where worldwide users publish and share their feelings. However, sentiment analysis for Twitter messages ('tweets') is regarded as a challenging problem because tweets are short and informal. In this paper, we apply a novel approach for automatically classifying the sentiment and emotions of Twitter messages. These messages are hierarchically categorized on basis of neutrality, polarity (positive or negative) and presence of various emotions. The hierarchical classification approach (HC) is a specialization of the well-known flat classification task. The main difference between them is that when using HC, examples must be assigned to classes organized in a previously defined class hierarchy; while traditional flat classification does not take into account the hierarchical information. We applied our model to posts collected from Twitter regarding the 2011 season of the Brazilian Soccer League. Our results show that the proposed method outperforms the corresponding flat approach in emotion classification.**

Keywords: hierarchical classification, sentiment and emotions classification, Twitter

I. INTRODUCTION

Twitter is a microblogging service where users post messages ("tweets") of no more than 140 characters. Twitter represents one of the largest and most dynamic real datasets of user generated and distributed content. These tweets sometimes express opinions and sentiment about different topics. With these data, companies have the opportunity to examine what customers are saying about their products and services.

Computational tools that automatically extract and analyze relevant information about opinion and sentiment expressed on Twitter and other social media sources are thus in high demand. Full sentiment analysis for a given specific topic or product requires many stages, including: (1) extracting tweets based on a specific query, (2) filtering out spam and irrelevant items from the data of those tweets, (3) identifying subjective tweets, and (4) identifying the polarity and the emotions of those tweets.

The simplest approaches are based on the presence of words or emoticons that are indicators of positive or negative polarity (e.g. Twitter's own API) [1], or calculating a ratio of positive to negative terms [2]. Though these are a useful first pass, the nuance of language often defeats them [3].

The standard (flat) supervised classification methods have been applied to recognition of positive or negative polarity [3-5], by using texts labeled with polarity as input. One way around this is to use noisy labels (also referred to as "distant supervision"), for example, by taking emoticons like ':)' as positive and ':(' as negative, and then training a standard classifier [6,7].

Semi-supervised methods can reduce dependence on labeled texts: for example, through the use of a polarity lexicon combined with label propagation [8]. Several have used label propagation, starting with a small number of hand-labeled words to induce a lexicon for use in polarity classification [7,9-11].

Our proposed supervised hierarchical classification is a novel approach in emotional analysis for Twitter messages (tweets), which considers the relation between the polarity and emotion of a tweet. The main idea is to order these categories and their relations in a hierarchical form and perform classification based on this hierarchy [12,13].

In contrast to flat classification, which builds models from a set of categories, which do not obey any order or structure, the main goal of hierarchical classification is to classify examples by following a predefined hierarchy.

This paper looks at the automatic recognition of not only the positive or negative polarity of a tweet, but also places it into one of seven emotion classes. The basic six classes are happiness, sadness, fear, anger, disgust, and surprise [14] and also a " no-emotional" class for the sentences, which bear no emotion. These seven emotion classes are used in previous work [12-17].

In this paper, we use a three-level hierarchical classification. The first step is to determine if the tweets (instances) are emotional. Next, the tweets defined as emotional in the previous step are classified based upon their polarity. In the third step we assume that, among the six emotions, the instances of happiness have positive polarity, while the other five emotions are regarded as having negative polarity. That is why we take the negative instances from the second step and classify them into the five negative emotion classes [12,13].

Our experiments on data annotated with emotions collected from Twitter surrounding the 2011 season of the Brazilian Soccer League show that this approach significantly outperforms the corresponding flat approach. We note significantly improved precision, recall and F-measure of all emotional classes.

The remainder of this paper is organized as follows. Section 2 briefly presents previous and related work. Section 3 gives an overview of the dataset and the experiment's feature sets. In Section 4 we describe the hierarchical classification method, and we evaluate it by comparing it to previous flat classification results. Finally, in Section 5 we present the conclusions and discuss the future works.

II. Related Works

There has been progress in research on sentiment analysis, however little work has been done regarding automatic recognition of emotion in tweets. Much work in sentiment analysis involves the use and generation of dictionaries which capture the sentiment of words [18]. These methods range from manual approaches of developing domain-dependent lexicons to semi-automated approaches and fully automated approaches [5,7,19].

Reference [1] uses the OpinionFinder subjectivity lexicon to label the polarity of tweets about Barack Obama and compare daily aggregate sentiment scores to the Gallup poll time series of manually gathered approval ratings of Obama. Even with this simple polarity determination, they find significant correlation between their predicted aggregate sentiment per day and the Gallup poll. Using the OMD dataset, in [20] Shamma et al. find that the amount of Twitter activity is a good predictor of topic changes during the debate, and that the content of concurrent tweets reflects a mix of the current debate topic and Twitter users' reactions to that topic.

Reference [21] performs aggregate sentiment analysis on tweets over time, comparing predicted sentiment to time series such as the stock market and crude oil prices, as well as major events such as Election Day and Thanksgiving.

The standard supervised classification methods have been applied to the recognition of positive or negative polarity [4], by using texts labeled with polarity as input. One way around this is to use noisy labels (also referred to as "distant supervision"), for example, by taking emoticons like ':)' as positive and ':(' as negative, and then training a standard classifier [6-7].

Reference [12] proposed a hierarchical classification approach to classifying blog sentences into Ekman's six emotion classes, as well as one proposed non-emotional class. They focus their work on emotional analysis and classification of emotions in text using blog datasets.

III. Proposed Approach and the Experimental Results

A. *The Data Sources and the classification process*

We applied our model to posts collected from Twitter regarding the 2011 season of the Brazilian Soccer League. To build the soccer dataset, we considered the 12 most popular Brazilian soccer teams, which play in the Brazilian 2011 First Division Soccer League. This data is collected and used by our application for opinion mining and sentiment analysis, called *SM-Brasileirao*. This application is used to analyze the reaction of fans during a soccer match in the Brazilian championship using the Twitter as the source (i.e. match between fans on Twitter before, during and after the games!). Figure 1 shows a snapshot of the *SM-Brasileirao*.

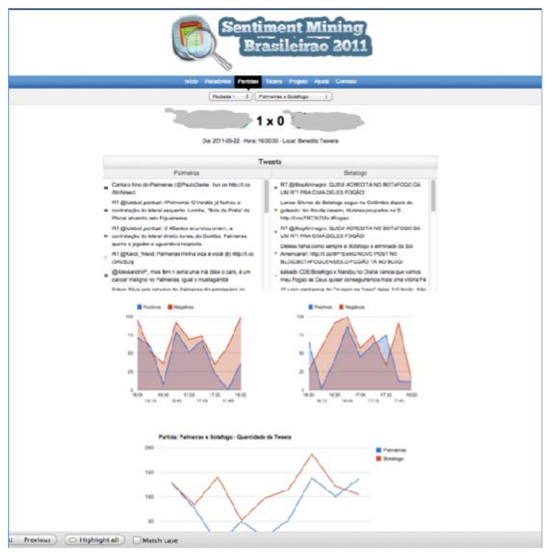

Figure 1. A snapshot of the *SM-Brasileirao*: Shows the polarity (positive and negative) emotion for Brazilian Soccer League: Available at: http://Licesa.dcc.ufla.br/brasileirao

The dataset contains 2809 annotated tweets collected during one month and manually annotated by three judges. Each tweet is tagged by a dominant emotion in the sentence, or as non-emotional if it does not include any emotion. To balance the dataset, as originally it had 67.1% non-emotional (the highest proportion) and the *surprise* class with only 2.8% of tweets, we removed 901 of the non-emotional tweets. Table I shows the specifications of the chosen dataset.

TABLE I. DATASET SPECIFICATIONS

Size	Num. of classes	Min-Max
1908	7	8.8-42.2%

In our experiments we used corpus-based feature sets, so we needed no other external resources, and we used the multiclass SVM as a machine-learning algorithm from Weka [22]. We also experimented using with the Naïve Bayes, however the SVM gives better results, as we get accuracy of 80.35% as a result of the emotional classification experiments and setting "10-fold cross validation" as a testing option.

B. *Three-level Hierarchy Classification and the Results*

In the past decade, research efforts in these areas have turned to the task of hierarchical classification. In contrast to flat classification, which builds models from a set of categories (which do not obey any order or structure), hierarchical classification has as its main goal to classify examples following a predefined hierarchy [23-25].

In this work, we use a three-level hierarchy classification approach, which addresses the relation between polarity and emotions, in addition to the relation between emotion and neutrality. This is based on the assumption that, among the six chosen emotions, happiness belongs to the positive polarity class, while the other five emotions are regarded as having negative polarity [12,25].

In this approach, we break the task of the emotion classification into three levels. The first level, emotional versus non-emotional classification, tries to determine whether a tweet is neutral or emotional. In the second level, we further classify tweets that have been identified as emotional as either positive or negative. We only consider happiness as positive. Finally, we classify the negative instances into five negative emotion classes. The results of this classification are shown in Table II.

As we can see in the results of Level 2, we increased the precision of the Positive (happiness) class. In the second level (polarity classification) the data are almost balanced, with 47% positive and 53% negative instances. That makes the instances defined as positive more precise.

By comparing the results of the three-level approach with the flat classification, we can see that the F-measure of all the emotional classes in the three-level experiment is higher than the F-measure of the emotional classes in the flat classification, with the exception of the "anger" class. In the anger emotion class, however, the F-measure of both is equal, but the precision of the three-level approach is higher. Figures 2, 3 and 4 show the comparison graphs of the F-measure, Precision and Recall of each class for the two approaches.

TABLE II. THE RESULT OF THREE-LEVEL VERSUS FLAT EMOTIONAL CLASSIFICATION (IN BOLD THE HIGHEST VALUES FOR EACH CLASS)

Hierarchical Three-level Classification					**Flat Classification**		
		Precision	*Recall*	*F-measure*	*Precision*	*Recall*	*F-measure*
Level 1	*emotion*	0.85	0.72	**0.78**			
	non-emotional	**0.77**	**0.88**	**0.82**	0.56	0.69	0.62
Level 2 (Based on result of Level 1)	*Positive*	**0.94**	**0.86**	**0.90**	0.74	0.44	0.55
	Negative	0.88	0.95	**0.91**			
Level 3 (Based on result of Level 2)	*sadness*	**0.57**	**0.43**	**0.49**	0.55	0.27	0.36
	fear	**0.74**	**0.86**	**0.80**	0.73	0.83	0.78
	surprise	**0.69**	0.91	**0.78**	0.54	**0.93**	0.68
	disgust	**0.55**	**0.37**	**0.44**	0.45	0.14	0.21
	anger	**0.61**	0.62	**0.62**	0.53	**0.73**	**0.62**

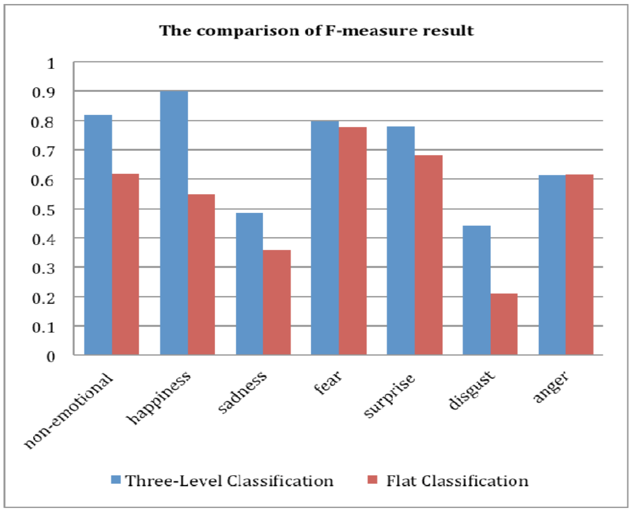

Figure 2. The comparison of F-measure result

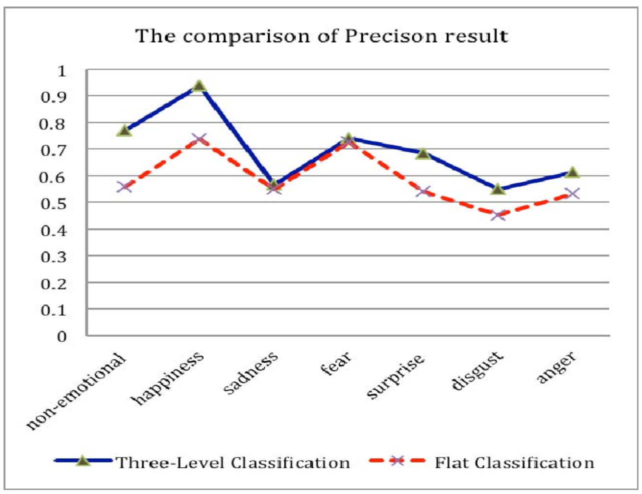

Figure 3. The comparison of Precison result

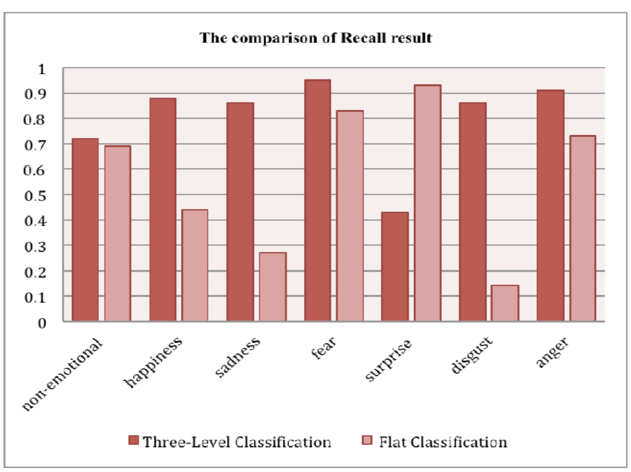

Figure 4. The comparison of Recall result

IV. CONCLUSIONS AND FUTURE WORKS

This paper presented a study on hierarchical classification to analyze emotion in tweets. We apply it to a specific dataset collected from Twitter on the Brazilian Soccer League. A manually annotated dataset is used to generate the corpus-based feature sets and is used by the SVM classifier to generate our experiments. We apply the flat classification approach and compare it with our three-level hierarchical classification. In our three-level approach, the first level is to identify the emotional versus non-emotional classification. In the second level, we further classify tweets that have been identified as emotional as either positive or negative. We consider only happiness as positive. Finally, we classify the negative tweets into five negative emotion classes. The results demonstrate a considerable improvement in the classification results.

In the future, we plan to expand our work by testing the approach on other Twitter datasets. Another direction for future work will be to expand our annotated corpus by using the emotional lexicon and their corresponding features. We believe that using the features derived from the emotional lexicons will improve the final results.

ACKNOWLEDGMENT

We would like to thank CNPq, FAPEMIG (Brazilian agencies) and NSERC (Canada) for partial financial support. The authors also thank the anonymous reviewers for useful remarks and suggestions.

REFERENCES

[1] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In Proceedings of the International AAAI Conference on Weblogs and Social Media , 2010.

[2] Yejin Choi and Claire Cardie. Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing , pages 590– 598. Association for Computational Linguistics, 2009.

[3] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval , 2(1-2):1–135, 2008.

[4] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing- Volume 10 , pages 79–86, 2002.

[5] 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, 2004.

[6] 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.

[7] Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph. In Proceedings of the EMNLP 2011 Workshop on Unsupervised Learning in NLP, pages 53-63, 2011.

[8] Vikas Sindhwani and Prem Melville. Document-word coregularization for semi-supervised sentiment analysis. In Proceedings of IEEE International Conference on Data Mining (ICDM-08) , 2008.

[9] [9] Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. Reis, and J. Reynar. Building a sentiment summarizer for local service reviews. In WWW Workshop on NLP in the Information Explosion Era (NLPIX) , 2008.

[10] Delip Rao and Deepak Ravichandran. Semi-supervised polarity lexicon induction. In Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009) , pages 675–682. Association for Computational Linguistics, 2009.

[11] Samuel Brody and Noemie Elhadad. An unsupervised aspect sentiment model for online reviews. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics , HLT '10, pages 804–812, Stroudsburg, PA, USA, 2010.

[12] Diman Ghazi, Diana Inkpen, and Stan Szpakowicz, "Hierarchical approach to emotion recognition and classification in texts", in Proceedings of the 23rd Canadian Conference on Artificial Intelligence (AI 2010), Ottawa, ON, Canada, May 2010, pp. 40-50

[13] Diman Ghazi, Diana Inkpen, and Stan Szpakowicz, "Hierarchical versus Flat Classification of Emotions in Text", in Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, Los Angeles, CA, June 2010, pp. 140-146.

[14] Ekman, P.: An Argument for Basic Emotions. Cognition and Emotion 6, 169–200 (1992)

[15] Neviarouskaya, A., Prendinger, H., Ishizuka, M.: Compositionality Principle in Recognition of Fine-Grained Emotions from Text. In: Proc. Third International ICWSM Conference, pp. 278–281,2009.

[16] Strapparava, C., Mihalcea, R.: SemEval-2007 Task 14: Affective Text (2007)

[17] Strapparava, C., Mihalcea, R.: Learning to Identify Emotions in Text. In: Proc. ACM Symposium on Applied computing, Fortaleza, Brazil, pp. 1556–1560 (2008)

[18] S. Das and M. Chan. Extracting market sentiment from stock message boards. Asia Pacific Finance Association, 2001.

[19] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining , pages 168–177, New York, NY, USA, 2004. ACM. ISBN 1-58113-888-1.

[20] Nicholas A. Diakopoulos and David A. Shamma. Characterizing debate performance via aggregated twitter sentiment. In Proceedings of the 28th international conference on Human factors in computing systems , pages 1195–1198, 2010.

[21] Bollen, A. Pepe, and H. Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In Proceedings of the 19th International World Wide Web Conference , 2010.

[22] Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)

[23] C. Silla and A. Freitas, "A survey of hierarchical classification across different application domains," Data Mining and Knowledge Discovery, vol. 22, pp. 31–72, 2011.

[24] Kiritchenko S., Matwin S., Famili F., "Hierarchical Text Categorization as a Tool of Associating Genes with Gene Ontology Codes", ECML/PKDD 2004 Workshop on Data Mining and Text Mining for Bioinformatics, Pisa, pp. 26-30, 2004.

[25] M. Ceci and D. Malerba, "Classifying web documents in a hierarchy of categories: a comprehensive study," J. Intell. Inf. Syst., vol. 28, no.1, pp.37–78, 2007.