Content and Style Features for Automatic Detection of Users' Intentions in Tweets

Helena Gómez-Adorno[1](✉), David Pinto[2], Manuel Montes[3], Grigori Sidorov[1], and Rodrigo Alfaro[4]

[1] Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico, Mexico
helena.adorno@gmail.com, sidorov@cic.ipn.mx
[2] Facultad de Ciencias de la Computación,Benemérita Universidad Autonóma de Puebla, Puebla, Mexico
dpinto@cs.buap.mx
[3] Coordinación de Ciencias Computacionales, Instituto de Astrofísica, Óptica y Electrónica, Puebla, Mexico
mmontesg@ccc.inaoep.mx
[4] Escuela de Ingeniería Informática,Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile
rodrigo.alfaro@ucv.cl

Abstract. The aim of this paper is to evaluate the use of content and style features in automatic classification of intentions of Tweets. For this we propose different style features and evaluate them using a machine learning approach. We found that although the style features by themselves are useful for the identification of the intentions of tweets, it is better to combine such features with the content ones. We present a set of experiments, where we achieved a 9.46 % of improvement on the overall performance of the classification with the combination of content and style features as compared with the content features.

Keywords: Short texts · Text classification · Twitter · Detection of intention

1 Introduction

Nowadays, social networks have become an important interaction media among worldwide users. Among the most used social networks is Twitter, a microblogging social network, with over 200 million users and about 400 million posts per day [1]. Twitter is used for various purposes by a large number of users, which may found themselves overwhelmed with the constantly growing amount

This work was done under partial support of the Mexican Government (CONACYT-134186, CONACYT grant #308719, SNI, COFAA-IPN, SIP-IPN 20144274) and FP7-PEOPLE-2010-IRSES: "Web Information Quality Evaluation Initiative (WIQ-EI)" European Commission project 269180.

A.L.C. Bazzan and K. Pichara (Eds.): IBERAMIA 2014, LNAI 8864, pp. 120–128, 2014.
DOI: 10.1007/978-3-319-12027-0_10

of received messages. In our personal experience, this is a major problem when the messages are accessed via mobile devices.

Some studies of Twitter identify that people use microblogging to talk about their daily activities and to seek or share information [2], making it a rich source for text analysis in several areas. Therefore, classification of tweets is an active research field. There are numerous research works on this social network in the area of sentiment analysis [3–5], predicting box office revenues [6] or the outcome of political elections [7], among others.

The classification of Twitter users intentions [8] is becoming a new opportunity area of research. Our aim is to identify intention or purpose of users typing a tweet. "Intention" is defined as an "agent's specific purpose in performing an action or series of actions, the end or goal that is aimed at" [9]. The automatic classification of tweets into intention categories may improve navigation and search for twitter users, especially when using a mobile device. In order to classify the user intention we use a taxonomy of 8 categories of the main user intentions in Twitter proposed in [10]. This taxonomy allows to classify messages in categories such as News Report (NR), News Opinion (NO), Publicity (PU), General Opinion (GO), Share Location/Event (SL), Chat (CH), Question (QU) and Personal Message (PM).

Most of the systems and approaches implemented to automatically detect the intention of the twitter message use a content based representation (raw word representation, user meta-information) as features to build a model for intention detection [10,11]. The contribution of this paper consists in analyzing the relevance of content and style attributes for this task. Based on tasks such as Authorship Attribution [12], Profiling [13] and Sentiment Analysis [5,14] we consider that the style plays an important role and complements the content information. The style information depends on the presence of pronouns, adjectives, verbal time, url and hashtags. We build a tweet representation extracting a set of content (words, words n-grams) and style (presence of hashtags, presence of emoticons, POS tags) features, which are subsequently used in a machine learning algorithm in order to built a classifier based on several labeled examples.

The obtained results show that the models created with the content and style features together overcome the results of the models using only words as features. Besides, the models created with the style features are domain independent, since they are no longer based on specific words but only on the language structure.

The rest of this paper is organized as follows. First we present an overview of related works on twitter text classification in Section 2. In Section 3 we explain our strategy for feature selection. In Section 4 the experimental setup is described and the obtained results are discussed. Finally, conclusions and future work directions are given in Section 5.

2 Related Work

There are a lot of research work in topic text classification (e.g. health, education, politics), hashtag recommendation, sentiment analysis; but there are a lack of studies where user intentions in Twitter are identified. In [2] the authors analyze the aggregate behavior across communities of users to describe a community intention. They also propose a user intention taxonomy based on the link structure of a community network. Even though, it does not propose an automatic methodology, this work laid the foundations of the task of detection of user intentions in Twitter.

In [15] the authors classify the tweets based on their content in 9 different twitter content type categories. They study in which way the tweets' content varies according to the users activity, personal networks and usage patterns. In the same way, in [8] the authors present a taxonomy of tweets purposes. Their aim was to identify user purposes in writing single tweets. Although, both works present an interesting taxonomy for detection of user intention in writing single tweets, none of them present an automatic method for classification.

With the purpose of improving of information filtering, in [11] the authors classify tweets in 5 general types of content (related to intention). They propose an automatic method using a set of domain-specific features extracted from the author's profile and text. In addition, they claim that corporate and personal Twitter users have different intentions. Such features, in comparison to the style features proposed here require external resources and user profile information.

In [10] the authors define a user intention taxonomy and an automatic classification model. They transform the tweets into a Vector Space Model, where the words are the features represented by the ("exists" or "does not exists") boolean values. Their results show that the tweets emitting an opinion were easily confused and consequently achieve low classification performance. There is a high overlap between two categories, thus leading to confusion of the classifier, given that the *News Opinion* tweets contain much of the content that the *News Report* expresses. By analyzing this phenomenon we propose the use of the style features which will be useful in this problem, since it has been proven to be helpful features in some related problems [11,16].

Another related problem is addressed in [16], where the authors use features related to certain content keywords in order to identify real-time intentions tweets. For this binary problem (tweet express or does not express an intention) the authors use a classification algorithm with the content keywords features, certain verbs, temporal expressions and POS tags with their position with respect to the content keywords. The POS features used in this task were useful, when they are employed in conjunction with other features.

3 Feature Selection

The aim of this work is to evaluate content and style features to identify user intentions when writing a single tweet. The features were selected taking into account the way in which each of them may represent certain classes. For example, Twitter users express themselves differently when sharing a *Personal Message* (PM) or *Chats* (CH) than when posting *News Reports* (NR) or *Publicity* (PU). In general, PM tweets tend to have mentions to other users and emoticons, while NR tweets are written in a clear form and usually using URLs to complement the news. In this sense a feature for discriminating PM tweets may be the use of mentions and emoticons, and viceversa the absence of such features will possibly discriminate NR tweets.

Let us consider the use of POS tags in tweets. For example, for the intention *Share Location/Event* (SL), we observe that the use of the verb in the gerund form is used frequently. Hence, we decided to label the tweets with their POS tags. The POS tags provide grammatical information of the words in the messages; in this way the classifier may be able to identify, say, SL messages by the presence of verbs in gerund forms among others. Interrogative pronouns are also identified by the POS tagger. Such words are useful for the identification of *Question Messages* (QU).

For the case of the opinion tweets (*News Opinion* and *General Opinion*), our hypothesis is that the presence of adjectives to express an opinion is very important. Therefore, these classes can be identified by the presence of such grammatical category. In the same way, function words (pronouns, prepositions, conjunctions, determiners, auxiliary verbs) have been effectively used for capturing author style in the task of authorship attribution [17]. We believe that function words tags may also be helpful features for the identification of user intentions in tweets.

Formally, tweets are represented as a vector $V = \{f_1, f_2, f_3, ..., f_n\}$, where n is the total number of features f_i. The set of features is divided into two subsets $V_1,\ V_2\ \subset\ V$ where V_1 represents the content features and V_2 represents the style features. An explanation of the details for each subset follows.

3.1 Subset of Content Features

The content features are represented by the words in the message, so we use bag-of-words and n-grams representations for these features. In order to assign weights to each feature, we have evaluated three different weight schemata: the term frequency (TF), the term frequency/inverse document frequency (TF/IDF), and the term presence (TP). Nevertheless, the performance of the classifier is practically the same for all types of values, probably due to the length of the tweet, which is 140 characters maximum. Hence, we decided to use the TP for the values of each feature; the presence of a word is identified with the value 1 and it's absence with the value 0.

3.2 Subset of Style Features

In this section we present the style features used in this work. For this purpose we extracted 4 style features, which were previously used in the related task of sentiment analysis [4]. For each tweet we identify: (a) initial mentions (presence of *@user* at the beginning of the tweet), (b) mention inside the message (presence of *@user* inside the tweet), (c) URL, and (d) Emoticons. The four previous features have binary values (BF, binary features), in the sequel 4BF, so the presence or absence of each one is identified with the values 1 or 0 respectively in the feature vector. Words and punctuation marks are also included in the features vector in the same way as the 4 previously explained features, indicating their presence and their absence with 1 and 0, respectively.

In addition, tweets are grammatically tagged using Freeling[1] for Spanish, obtaining the Part-Of-Speech (POS) tags of each word in the tweet. Freeling uses the EAGLES[2] Standard in order to represent the grammatical information of words. Each label is composed of at most 8 digits, where the first digit represents the grammatical category and the rest of the digits represent the attributes of each grammatical category. The total number of POS tags used was 240. The POS tags become additional features in the feature set, and the presence or absence of each one is indicated with the values 1 or 0 in the feature vectors, respectively.

4 Experimental Results

In this section we describe the experimental setup and the results obtained when evaluating the different features considered in our approach.

4.1 Experimental Setup

The dataset used in this study is the one described in [10]. The dataset consists of 5,209 messages manually classified by the authors of [10] into the following classes: News Report (NR), News Opinion (NO), Publicity (PU), General Opinion (GO), Share Location/Event (SL), Chat (CH), Question (QU) and Personal Message (PM). The class distribution given in Figure 1. The tweets are written in Spanish and contain mentions referring to: *Banco Santander* (a bank), *Pontificia Universidad Católica de Valparaíso* (a university), *El Mercurio* (a newspaper), *La Tercera* (a newspaper), *Movistar* (a telecommunication company) and a set of random tweets.

We used seven different configurations for our test dataset. First, we evaluated the classification performance when using only the content features (only words unigrams or words + word bigrams + word trigrams). Second, the classification performance was evaluated with the style features (only 4BF or POS + 4BF or POS + POS bigrams + POS trigrams + 4BF). Finally, we have mixed the content and style features, obtaining two different feature sets: words + POS + 4BF or Word_POS + Word_POS bigrams + Word_POS trigrams + 4BF. So, each feature type was evaluated independently in order to determine which one is most suitable for this task. Thereafter, the combination of the types of features was evaluated in order to assess whether they complement each other to improve the classification accuracy.

For the classification process we used a machine learning approach. The experiments were conducted using the SVM algorithm provided in the WEKA[3] data mining tool. The training phase was performed on the entire data set using

[1] http://nlp.lsi.upc.edu/freeling/
[2] http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html
[3] http://www.cs.waikato.ac.nz/ml/weka/

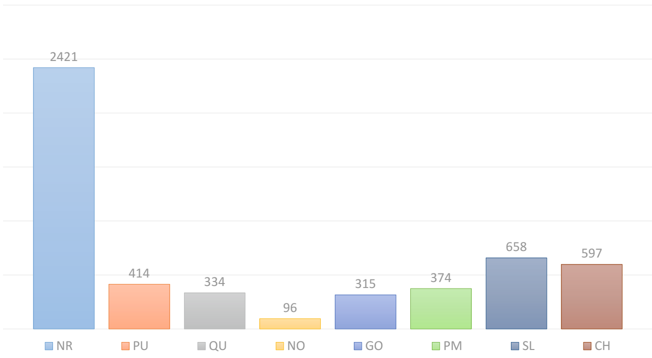

Fig. 1. Class distribution in the dataset.

tenfold cross-validation. The evaluation metrics used were the $F1$ and $F1_\mu$ measures :

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{1}$$

$$F1_\mu = 2 \cdot \frac{precision_\mu \cdot recall_\mu}{precision_\mu + recall_\mu} \tag{2}$$

where:
tp =number of true positives (correct class prediction),
tn =number of true negatives (correct negative prediction),
fp =number of false positives (incorrect class prediction),
fn =number of false negatives (incorrect negative prediction),
i = number of classes,

$$precision = \frac{tp}{tp + fp}, \tag{3}$$

$$recall = \frac{tp}{tp + fn}, \tag{4}$$

$$precision_\mu = \frac{\sum_{i=1}^{|c|} tp_i}{\sum_{i=1}^{|c|} tp_i + fp_i}, \tag{5}$$

$$recall_\mu = \frac{\sum_{i=1}^{|c|} tp_i}{\sum_{i=1}^{|c|} tp_i + fn_i}. \tag{6}$$

4.2 Evaluation

In this section we present the evaluation of the classification using the above-mentioned seven feature sets. With this experiment we aimed to assess the benefits of using style features such as POS and POS n-grams over content features such as words and word n-grams.

As mentioned in section 3, in Figure 2 the label 4BF refers to the four features previously presented: initial mention, mention inside the message, URL, emoticons. The labels "words and words n-gram" refer to the entire set of words of the corpus and the n-gram of words, with $n = 1, 2, 3$. The labels "POS and POS n-grams" refer to the set of POS tags corresponding to the words and the n-grams of such POS tags with $n = 1, 2, 3$. Finally, the label "(Word_POS) n-grams" refers to the combination of a word and its POS tag as one feature.

The results presented in Figure 2 depict the $F1_{\mu}$ measure of the classification over the data sets. The Figure shows that the use of the 4BF by themselves does not outperform the performance of the classification when using the content features. On the contrary, POS + 4BF features improves the performance by 6.49% over the words' features. Besides, the combination of a word with its POS tag plus the 4BF achieves a 9.46% of improvement over the Words features.

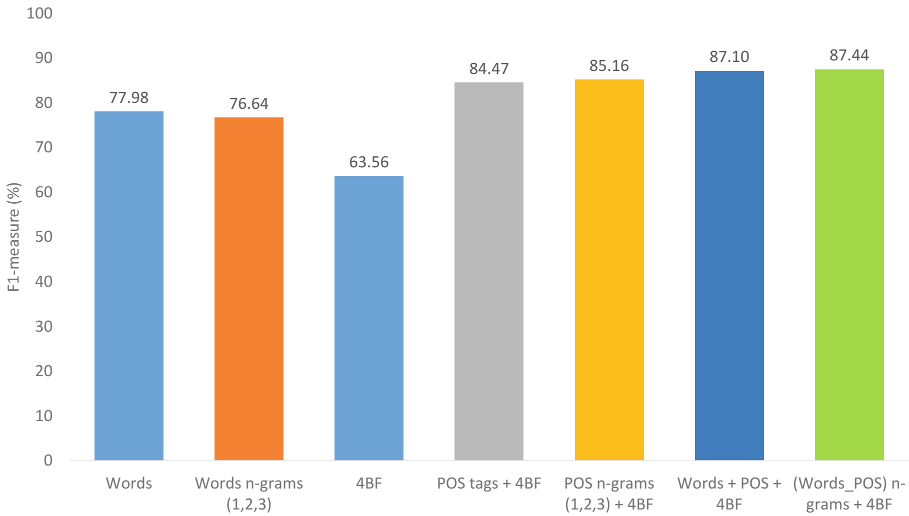

Fig. 2. Classification performance obtained with different feature sets

Table 1 presents the detailed results, in $F1$ measure terms, for each type of feature sets per class. We can observe that the 4BF by themselves do not identify the tweets in the classes *Publicity*, *Question*, *News Opinion* and *General Opinion*. On the other hand, such features are able to identify the classes *News Report* and *Chat* with 81.4% and 68.1% of $F1$ measures, respectively. In addition, the style features (POS n-grams + 4BF) outperform the content features (words and word n-grams) in almost all classes, with the exception of the *General Opinion* and *Publicity* classes. The class that benefits the most from the style features is *Question*, achieving an $F1$ measure of 87.5% with such features and only 53.6% with the content features. Other classes such as *News Report*, *Personal Messages*, *Share Location* and *Chat* also benefit from the style features. On the contrary, such classes as *Publicity*, *News Opinion*, and *General Opinion* need both the content features and the style features in order to improve the classification performance.

In Table 1, is observed that the use of POS features for the classes *Publicity*, *News Opinion*, and *General Opinion* did not result in an improvement of the classification. It is possible that the tagger does not identify well the grammatical tags due to the use of an informal language in tweets. These phenomena may bring as a result a decrease of the classification accuracy.

Table 1. $F1$ measures obtained with the different configuration of the feature set

Features	NR	PU	QU	NO	GO	PM	SL	CH
Words	88.6	80.0	53.6	49.7	39.4	48.6	94.8	62.9
Word n-grams	87.8	79.5	41.5	40.9	22.6	48.5	93.7	62.0
4BF	81.4	0.00	0.00	0.00	0.00	58.1	24.2	68.1
POS tags + 4BF	93.4	71.7	87.6	48.7	34.3	62.4	95.7	80.3
POS tag n-grams + 4BF	94.4	74.9	87.6	40.0	44.0	63.5	96.7	80.3
Words + POS tags + 4BF	95.4	80.4	**88.0**	53.6	**48.7**	**68.4**	96.9	81.4
(Word_POS) n-grams + 4BF	**95.6**	**80.7**	**88.0**	**55.6**	46.3	**68.4**	**97.1**	**82.9**

5 Conclusion and Future Work

We have presented an approach for automatic classification of tweets taking into account user intention categories. We used style and content features of tweets. This kind of classification can be used for filtering out tweets by user interests and facilitate queries and navigation.

Experimental results show that the POS features improve the performance of the classification as compared with only the words features. Besides, the combination of the word with its POS tag performs significantly better with this set of classes.

We are currently working on the selection of new features that could improve the classification of certain classes such as *News Opinion*, *General Opinion* and *Personal Message*. For example, we consider other style features, such as syntactic n-grams, that were successfully used in the task of authorship attribution and author profiling [18].

References

1. Wickre, K.: Celebrating #Twitter7. Twitter Blog (2013), https://blog.twitter.com/2013/celebrating-twitter7
2. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: Understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis. ACM (2007)

3. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of Twitter data. In: Proceedings of the Workshop on Languages in Social Media. ACL, Stroudsburg (2011)
4. Pandey, V., Iyer, C.: Sentiment analysis of microblogs. Technical report (2009)
5. Jain, H., Mogadala, A., Varma, V.: Sielers : Feature analysis and polarity classification of expressions from Twitter and SMS data. In: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). Second Joint Conference on Lexical and Computational Semantics (*SEM), vol. 2. ACL (2013)
6. Asur, S., Huberman, B.A.: Predicting the future with social media. In: Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 1. IEEE Computer Society (2010)
7. Diakopoulos, N.A., Shamma, D.A.: Characterizing debate performance via aggregated Twitter sentiment. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM (2010)
8. Alhadi, A.C., Staab, S., Gottron, T.: Exploring user purpose writing single tweets. In: WebSci 2011: Proceedings of the 3rd International Conference on Web Science. ACM (2011)
9. Website: Wikipedia, the free encyclopedia (2014), http://en.wikipedia.org/wiki/Intention
10. Martis, M., Alfaro, R.: Clasificación automática de la intención del usuario en mensajes de Twitter. In: I Workshop en Procesamiento Automatizado de Textos y Corpora, Sausalito, Viña del (March 2012)
11. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., Demirbas, M.: 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. ACM (2010)
12. Stamatatos, E.: A survey of modern authorship attribution methods. J. Am. Soc. Inf. Sci. Technol. 60(3) (2009)
13. Rangel, F., Rosso, P.: Use of language and author profiling: Identification of gender and age. Natural Language Processing and Cognitive Science (2013)
14. Barbosa, L., Feng, J.: Robust sentiment detection on Twitter from biased and noisy data. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters. ACL, pp. 36–44 (2010)
15. Naaman, M., Boase, J., Lai, C.H.: Is it really about me?: Message content in social awareness streams. In: Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work. ACM (2010)
16. Banerjee, N., Chakraborty, D., Joshi, A., Mittal, S., Rai, A., Ravindran, B.: Towards analyzing micro-blogs for detection and classification of real-time intentions. In: Breslin, J.G., Ellison, N.B., Shanahan, J.G., Tufekci, Z., (eds.) Proceedings of International AAAI Conference on Weblogs and Social Media. The AAAI Press (2012)
17. Zhao, Y., Zobel, J.: Effective and Scalable Authorship Attribution Using Function Words. In: Lee, G.G., Yamada, A., Meng, H., Myaeng, S.-H. (eds.) AIRS 2005. LNCS, vol. 3689, pp. 174–189. Springer, Heidelberg (2005)
18. Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., Chanona-Hernández, L.: Syntactic n-grams as machine learning features for natural language processing. Expert Syst. Appl. 41(3) (2014)