Mining Tweet Data - Statistic and semantic information for political tweet classification

Guillaume Tisserant, Mathieu Roche, Violaine Prince

Mining Tweet Data
Statistic and Semantic Information for Political Tweet Classification

Guillaume Tisserant[1], Mathieu Roche[1,2] and Violaine Prince[1]
[1]*LIRMM, CNRS, Université Montpellier 2, 161 Rue Ada, 34090 Montpellier, France*
[2]*TETIS, Cirad, Irstea, AgroParisTech, 500 rue Jean-Franccois Breton, 34093 Montpellier Cedex 5, France*

Keywords: Text Mining, Classification, Tweets.

Abstract: This paper deals with the quality of textual features in messages in order to classify tweets. The aim of our study is to show how improving the representation of textual data affects the performance of learning algorithms. We will first introduce our method GENDESC. It generalizes less relevant words for tweet classification. Secondly, we compare and discuss the types of textual features given by different approaches. More precisely we discuss the semantic specificity of textual features, e.g. Named Entities, HashTags.

1 INTRODUCTION

This paper deals with detection of *important* words in a document, and how to use them for classification. Importance is a notion that is not predefined. It depends on the task goal and features, as well as on the user's intentions. Textual data are extremely difficult to analyze and classify, according to (Witten and Frank, 2005). The supervised learning algorithms require to know the class associated with each document (e.g. theme for document classification, polarity for sentiment analysis, and so forth). The inputs of these algorithms are "package" of language features representing the document to be classified. Once the learning phase is complete, the trained model can assign a class to a " package of features " unlabeled. The quality of the classification given by the algorithm will therefore depend not only on the quality of the learning algorithm, but also on how the transmitted data is represented (Guyon and Elisseeff, 2003).

To sketch an overview of the question to be tackled, in the first instance, we run a brief survey of the different research methods description of textual data suitable for supervised learning. Then, we present GENDESC, a statistical method to select features to the purpose of text classification (Section 3). In Section 4, we compare information given by GENDESC with information extracted of tweets with "semantic" methods. Every new step needs to be confirmed: Thus, the GENDESC method is evaluated in Section 5, focusing on the quality of its proposed features. Finally, we draw the current balance of our work and present some perspectives in Section 6.

2 GENERALISATION OF TEXTUAL FEATURES

In the abundant literature about text mining, the traditional method of representation of textual data is the "bag of words" model: Words are considered as features used as inputs in the learning algorithms (Salton and McGill, 1986). Despite its popularity, this method has many limitations. First, it highlights a large number of features : The matrix of features is larger than the number of terms appearing in the corpus, even though short texts like tweets do not provide many features per document (Sriram et al., 2010a). Secondly, it loses all the information related to the position of the words and their syntactic roles in the sentence, as well as all the information related to the context. This information can be sometimes crucial.

Several researchers have stressed the importance of having more general features than words. Experiments have shown the possibility of using the radical or the lemma of each word (Porter, 1980), or the *POS-Tag* (Gamon, 2004). But these methods are not satisfactory. Lemma is not an important generalization, as generalizing words with lemmas cannot significantly decrease the number of features. POS-TAG is an important generalization which destroys all semantic information. It can be useful for a specific task like grammatical correction, but is unadapted for ev-

Tisserant G., Roche M. and Prince V..
Mining Tweet Data - Statistic and Semantic Information for Political Tweet Classification.
DOI: 10.5220/0005170205230529
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (SSTM-2014), pages 523-529
ISBN: 978-989-758-048-2

ery word in a generic text classification task. For purposes of classification on specific tasks, some studies like (Joshi and Penstein-Rosé, 2009) propose to generalize specific words (determined according to their grammatical role) by substituting a more general feature and keeping the other in their natural form. In this kind of method, words are selected to be generalized using information based on their classification type. GENDESC, our method, is independent of the classification task. Its goal is to improve the quality of the classification of textual data, when deprived of any knowledge on the classification task.

3 GENDESC APPROACH

The approach we propose is in two steps. The first one determines the grammatical function of each word in the corpus (Section 3.1). The second step consists in selecting the words which will be generalized (Section 3.2).

3.1 Partial Generalization

The objective of our work is to generalize words whose POS tag (in its context of appearance) is more useful than the word itself for the classification task. A POS tag (i.e. verb, noun, adjective, etc.) is a class of objects that behave similarly when it comes to the sentence building. So, replacing a word by its tag comes to delete a semantic contents and keep the grammatical role. This task needs to explore a few tracks to determine the words that must be generalized. We start by replacing the words that appear in few documents by their POS tag. Those that appear often are deemed relevant for classification and are directly used as features.

Take for example the corpus of Figure 1. It contains 14 documents, each consisting of a single sentence describing a *semantic relationship type* (h: hyperonym; s: synonym). In this context, the aim of classification is to predict the relationship associated to each document (here, sentence). From the example in Figure 1, we obtain, for each word, the number of documents containing it. Actually, the example is built with short sentences like tweets.

Then, for each word of the sentence, the POS-tag is computed.

• **Example of Sentence:**
The molecular data is also sometimes called a gene clock or evolutionary clock.

This practice of adding minerals to herbal medicine is known as rasa shastra. (s)
The ellipsis is called wielokropek. (s)
The pancreas is a sort of storage depot for digestive enzymes. (h)
Walnuts and hickories belong to the family Juglandaceae. (h)
In both group the anterior tagma is called the ceprhalothorax. (s)
Biochemistery, sometimes called biological chemistery, is the study of chemical processes in living organisms. (s)
Inhalational anthrax is also known as Wooksorters' or Ragpickers' disease.(s)
Biology is a natural science concerned with the study of life and living organisms. (h)
This part of biochemistry is often called molecular biology. (h)
Biological classification belongs to the science of zoological systematics. (h)
The philosophy of biology is a subfield of philosophy of science. (h)
An MT is also known as a Medical Language Specialist or MLS. (s)
The molecular data is also sometimes called a gene clock or evolutionary clock. (s)
Stotting, for instance, is a sort of hopping that certain gazelles do when they sight a predator. (h)

Figure 1: Semantic relationship corpus

Table 1: number of instance of each word.

Word	Amount of documents containing the word
is	12
the	9
of	8
a	6
called	5
biology, science	
as, know, or, also	3
study, this, molecular	
for, sort	2

• **Example of the Previous Sentence Tagged with PoS:**
The/DT molecular/JJ data/NNS is/VBZ also/RB sometimes/RB called/VBN a/DT gene/NN clock/NN or/CC evolutionary/JJ clock/NN ./.

In this example, the POS-tags are:
RB: Adverb;
DT: Determiner ;
VB: Verb ;
VBN: Verb, past participle ;
VBZ: Verb, 3rd ps. sing. present ;
NN: Noun ;
JJ: Adjective ;

Table 2: Generalisation.

word	POS-tag	Number of documents which contain the word	feature
The	/DT	9	The
molecular	/JJ	2	/JJ
data	/NNS	1	/NNS
is	/VBZ	12	is
also	/RB	3	/RB
sometimes	/RB	1	/RB
called	/VBN	5	called
a	/DT	6	a
gene	/NN	1	/NN
clock	/NN	1	/NN
or	/CC	3	/CC
evolutionary	/JJ	1	/JJ
clock	/NN	1	/NN

To generalize some words of a sentence, we have tested some ranking functions which can assign a value to each word. If this value is smaller than a threshold, the word is generalized. The function used in the first example is the number of documents of the corpus that contains the word. So the threshold represents a minimum number of documents that must contain the word.

According to our example, if we consider 4 as a threshold, all the words that appear in less than 4 documents will be replaced by their POS-tag found at the previous step (see Table 2).

The initial sentence of our example, with this ranking function and a threshold at 4, is generalized as follows: *The /JJ /NNS is /RB /RB called a /NN /NN /CC /JJ /NN .*

We show below the same example with all possible generalization thresholds.

- threshold at 1 (we do not replace any word): *The molecular data is also sometimes called a gene clock or evolutionary clock.*
- threshold at 2: *The molecular /NNS is also /RB called a /NN /NN or /JJ /NN .*
- threshold at 3: *The /JJ /NNS is also /RB called a /NN /NN or /JJ /NN .*
- threshold at 4: *The /JJ /NNS is /RB /RB called a /NN /NN /CC /JJ /NN .*
- threshold at 6: *The /JJ /NNS is /RB /RB /VBN a /NN /NN /CC /JJ /NN .*
- threshold at 7: *The /JJ /NNS is /RB /RB /VBN /DT /NN /NN /CC /JJ /NN .*
- threshold at 10: */DT /JJ /NNS is /RB /RB /VBN /DT /NN /NN /CC /JJ /NN .*
- threshold at 13 (all words are replaced by their POS-tag): */DT /JJ /NNS /VBZ /RB /RB /VBN /DT /NN /NN /CC /JJ /NN .*

We can observe that, before the threshold becomes too important, the relevant information (for the relationship type given by the sentence) becomes more evident than in the full sentence. The less relevant words are the first to be generalized, and the words that give the semantic type of the relation ("*is*" and "*called*") are generalized only when the threshold is high. From threshold 4 to 6, a "definition" pattern appears : "The X is called a Y". This pattern is a strong clue for a semantic relationship between X and Y, and thus helps classifying the document. Thus, the question of the threshold value relevance appears: In section 5 we discuss the definition of different thresholds. Note we can combine this method with a filter to delete stop-words.

3.2 Ranking Function

In this example, we have tested the *DF* (Document Frequency) function, to generalize specific words. The more present text containing the word in the corpus, the higher its DF. We have tested some other ranking functions. For each function, we have tested several thresholds which will be discussed in Section 5.

3.2.1 IDF: Inverse Document Frequency

The *IDF* function (formula (1)) is the inverse function of *DF* (Jones, 1972).

$$IDF(x) = \log \frac{\text{number of documents in the corpus}}{DF(x)} \quad (1)$$

The more present a text containing the word is in the corpus, the lower its *IDF* is. Thus, the words with high *DF* value are those with a low *IDF* value. So, words generalized with a low threshold become those requiring a high threshold to be generalized. IDF function generalizes common words that appear in classes, which are therefore not interesting for classification.

For example, the *IDF* of the word *is*, which appears in many documents (i.e 12 sentences of the corpus of Figure 1) is $\log(\frac{14}{12}) = 0.07$ while *IDF* of word *for* is $\log(\frac{14}{2}) = 0.8$

3.2.2 TF: Term Frequency

TF is the term frequency in a document (Luhn, 1957). The more often a term appears in the document, the

lower its probability of being generalized. This function can be interesting because if a word appears many times in the same document, the word can be useful to classify documents.

3.2.3 D: Discriminence

The idea of D measure (formula (2)) is that a word which appears many times in the same class and never in other classes, is relevant for classification. For example, for semantic relationship classification, if a word appears often in documents that contain synonymy relationship, and never in other documents, then this word is probably relevant for synonymy relationship identification. D measure (see formula (2)) corresponds to the computation of $TF \times IDF$ where all documents belonging to the same class are considered as a single document.

$$D(x) = \frac{\text{nbOccClass(x)}}{\text{nbOccCorpus(x)}} \quad (2)$$

- $nbOccClass(x)$ is the number of occurrences of word x in the class that most often contain x
- $nbOccCorpus(x)$ is the number of occurrences of x in the entire corpus

3.2.4 Combination of Functions

It is possible to combine some functions. For example, the combination of the functions D and DF highlights words that both appear often and indicate a particular class.

We can expect than the words with high $DF \times D$ value are relevant for document classification.

Similarly, all functions described above can be combined together. For example, the function $D \times IDF$ emphasizes the words that appear rarely and are distributed very unevenly between classes. TF can be combined with all other functions. We tested these different functions with different thresholds. The efficiency of each function will be discussed in Section 5.

4 ANALYSIS OF FEATURES GIVEN BY GENDESC

4.1 Linguistic Features and Semantic Information

Semantic information is very difficult to use in text classification context, since semantic disambiguation is an crucial requirement before feeding semantic data to an automatic process: The meaning of a word is generally complicated to take into account for learning algorithms.

But it is the main information used by humans when they analyze documents. So it is probably the most important information to use in text processing.

In this paper, we focus on the semantic "importance" of a word in a document. The following subsections highlight *endogenous* (see Section 4.1.1) and *exogenous* (see Section 4.1.2) semantic information.

4.1.1 Endogenous Semantic Information

An endogenous method aims at using information contained in documents. Different approaches allow the detection of words with high semantic information (Faure and Nedellec, 1999; Hirano et al., 2007). Some methods consist of exploiting syntactic information in order to induce semantic information (Faure and Nedellec, 1999; Béchet et al., 2014). For instance, all the objects of the verb *to eat* can be gathered in a same concept *food*.

Other approaches use meta data, e.g. labels of HTML pages. For instance, the *title* and *keyword* labels can highlight significant features.

In social networks (e.g. Twitter) *HashTags* represent a semantic information useful for tweet classification (Conover et al., 2011; Ozdikis et al., 2012; Costa et al., 2013). *HashTags* are words highlighted by people who write tweets. This information can be considered as keywords. As an example, in 2012, the HashTag #2012 was used by many people to precise their tweet concern the presidential election. Other people use name of their favorite candidate to highlight their political opinion. Thus, the HashTag #Obama has been massively used to mark a support of Barack Obama. Supporters of the republican party have used #RomneyRyan2012, as a reference to the candidature of Mitt Romney for the presidential position and of Paul Ryan for the vice presidential role. Some HashTags are less explicit, like #Forward2012, exploited by the Obama campaign.

4.1.2 Exogenous Semantic Information

Exogenous methods rely on external sources of information to improve analysis of documents (Sriram et al., 2010b). There is a lot of documents representing semantic information, e.g., ontologies, lexicons, semantic networks, and so on. We use the lexical-semantic network called *JeuxDeMots* (JDM) (Chamberlain et al., 2013). It is based on the serious game principle in order to construct a large lexical-semantic

network. JDM is the most important semantic network for the French language. It is totally built by players.

Moreover, JDM contains a game named *Polititt* to generate political orientation of words. The construction of this game is very simple: Words are presented with six political orientations (i.e. far left, ecologist, left, center, right, far right). The user has to click on the orientations having most connections with the word. The user can skip the word if he considers it without semantic connection with a political orientation.

4.2 Comparison of Approaches

We argue that words considered relevant by a statistical analysis must contain important semantic aspects. So the aim of this section is to compare features returned by GENDESC (i.e. statistical method) with approaches using semantic characteristics. More precisely, we compare words given with *D* formula, HastTag, and *PolitIt*.

We can see in Figure 2 that a high proportion of HashTag words are considered relevant with GENDESC. In addition, more than 50% of the words of *PolitIt* that appear in the corpus are also considered relevant with GENDESC.

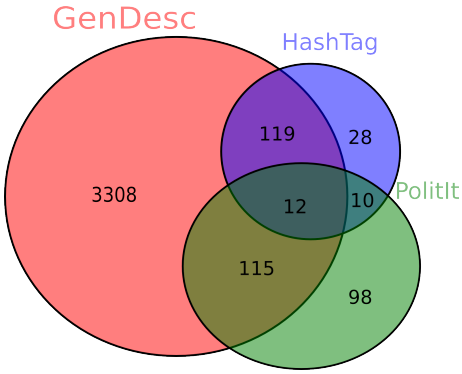

Figure 2: Venn diagram showing number of common words in the different information sources.

Experiments of Section 5.2 describe different types of words (e.g. Person, Location, and so on) given with the methods (i.e. GENDESC, HashTag, *PolitIt*).

5 EXPERIMENTS

5.1 Classification with GENDESC

5.1.1 Experimental Protocole

We chose to tackle a problem of classification in order to predict a political orientation of a tweet. We have used a corpus composed of 1500 tweets distributed between five political parties. The goal is to classify tweets based on the political party of the user.

In order to check the effectiveness of GENDESC, we tested three different learning algorithms: Bayesian classification algorithm, decision trees, and an algorithm based on SVM (Support Vector Machine). This point will be discussed at the end of this section 5.

Different functions have been tested to choose words which must be generalized. We can see that the quality of a function is almost independent of the learning algorithm.

Table 3 shows results obtained using NaiveBayes algorithm with different functions and different generalization thresholds. We used as baseline a classification with bag-of-words as features, obtaining an accuracy of **46.80%**.

Function and Threshold

Table 3 shows that only *D* function is actually relevant for the choice of words which can be generalized. Other functions generally provide lower results than use of bag-of-words such as features, whatever the threshold. For the *D* function, the optimal threshold is 0.3.

Table 3: Accuracy according to measures and thresholds.

Threshold	0.1	0.3	0.5	0.7
D	46.80	**50.24**	**49.19**	**47.38**
DF	46.32	45.44	43.90	43.70
IDF	**47.47**	41.95	26.85	25.10
TF	19.62	19.62	19.62	19.62
$DF \times D$	46.32	44.37	43.36	43.22
$D \times IDF$	**49.40**	35.50	20.13	20.13
$TF \times D$	46.80	46.26	43.22	45.37
$TF \times IDF$	39.66	28.19	22.80	21.95
$TF \times DF$	46.26	45.44	43.83	43.63

Our previous work (Tisserant et al., 2013) confirmed the interest of our method for classification tasks with different corpora.

Machine Learning Algorithms

Three algorithms were tested in their version implemented in Weka (Hall et al., 2009) :

- The Bayesian algorithm is NaiveBayes
- The decision tree is C4.5
- The algorithm based on SVM applied is SMO (Sequential minimum optimization)

These algorithms are used with default parameters of Weka using cross-validation (10-fold).

Experiments with several learning algorithms were run, in order to compare their performance. Table 4 shows the obtained results. SMO has the best performance and the algorithm based on decision trees has the lowest performance, whether using words as features or those obtained with GENDESC.

Table 4: Results obtained with different learning algorithms.

Method	GENDESC	**Bag-of-words**
SMO	55.30	52.33
NaiveBayes	50.24	46.80
C4.5	38.52	43.43

5.1.2 HashTag Generation

Another attractive issue of GENDESC concerns the HashTag generation (Mazzia and Juett, 2011). Since HashTags are neither registered nor controlled by any user or group, it is difficult for some users to determine appropriate HashTags for their tweets (Kywe et al., 2012). We have first compared the type of words given by GENDESC with the HashTag of the corpus, and then, we investigate if words selected with GENDESC in the corpus can be used as possible HashTags.

5.2 Types of Features Given with GENDESC

This section describes the types of features (e.g., named entities and HashTags) extracted with different systems.

Table 5 has been built with 25 words of each category manually annotated. It shows the different types of words returned with our systems. For example, data given by *PolitIt* contain a lot of political organisations.

We have compared the set of named entities given by the systems with the similarity measure *Cosine*. Table 6 presents the obtained results.

Table 5: Word classification in % with :
- Word with highest *D* value by using GENDESC
- Word most referenced in *PolitIt*
- HashTag most frequent in the corpus

Approaches	GENDESC	**PolitIt**	**HashTag**
Person	16	28	4
Location	28	16	12
Political Organisation	4	28	8
Non political Organisation	8	4	20
Other	44	22	56

Table 6: Similarity of different group of words.

Cosine(GENDESC, Politit)	0.73
Cosine(GENDESC, HashTag)	0.90
Cosine(HashTag, Politit)	0.59

Towards HashTag Generation

We show that words selected by people as HashTags contain important semantic information. Note that HashTag generation is a totally different task than classification; This enables to identify semantic information in messages.

Table 6 and Figure 2 highlight that ungeneralized words with GENDESC are close to HashTags. So we argue these words could be interesting as Hashtags. To validate this hypothesis, we have studied if words with high *D* values are currently used in Twitter as HashTags.

Table 7: Connexions between hashtags.org, PolitIt, and GENDESC.

GENDESC	PolitIt	GENDESC ∩ PolitIt
52%	76%	92%

Table 7 has been built thanks to hashtags.org website [1]. We have taken 25 words for each category in order to study how many are used more than a hundred time per day on average. Words selected with GENDESC are words with high *D* value. PolitIt words are those which have been attached with one political party by most of players. Measures have been done in July 2014, more than two years after the corpus acquisition.

This table shows that more than 50% of words with high value of *D* are used as Hashtags. However, the words that come from PolitIt are more relevant for Hashtag generation because of the specificities of the application. Note that the intersection of GENDESC and PolitIt gets a better result than using PolitIt words.

[1] http://www.hashtags.org

6 CONCLUSION AND FUTURE WORK

In this paper, we proposed a representation of textual data that improves classification of document methods by generalizing some features (words) to their POS category when these words appear as less discriminant for the task. Our results show that this approach, called GENDESC is appropriate when classification is at stake, regardless of the nature of its criteria. We have also demonstrated that *D*, *Discriminence*, is a measure that can be relevant to find semantically important words in a corpus. In our future work, we plan to use semantic information to improve classification.

In previous work, we proved that *n*-grams can be combined with GENDESC to slightly improve the classification (Tisserant et al., 2013). HashTag can probably be generated with *n*-grams of words with high *D* value. So we plan to use these *n*-grams in order to construct new Hashtags (e.g. *kdir 2014* $\rightarrow$ *#kdir2014*). They could be useful to detect HashTags which combine several concepts associated with *n*-grams returned with GENDESC (i.e. *n*-grams of words and/or Hashtags). As an example, a lot of tweets contain both HashTag *#Iran* and word *nuclear*, and they are often close to each other. The system should detect that *#IranNuclear* could be an interesting HashTag for all these tweets, which evoke the Iranian nuclear issue. If enough people use the proposed HashTag, they could follow news about "Iranian nuclear" more easily.

REFERENCES

Béchet, N., Chauché, J., Prince, V., and Roche, M. (2014). How to combine text-mining methods to validate induced verb-object relations? *Comput. Sci. Inf. Syst.*, 11(1):133–155.

Chamberlain, J., Fort, K., Kruschwitz, U., Lafourcade, M., and Poesio, M. (2013). Using games to create language resources: Successes and limitations of the approach. In *The Peoples Web Meets NLP*, pages 3–44. Springer.

Conover, M., Gonçalves, B., Ratkiewicz, J., Flammini, A., and Menczer, F. (2011). Predicting the political alignment of twitter users. In *Proceedings of 3rd IEEE Conference on Social Computing (SocialCom)*.

Costa, J., Silva, C., Antunes, M., and Ribeiro, B. (2013). Defining semantic meta-hashtags for twitter classification. In *Adaptive and Natural Computing Algorithms*, pages 226–235. Springer.

Faure, D. and Nedellec, C. (1999). Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system asium. In *In Proceedings of EKAW*, pages 329–334.

Gamon, M. (2004). Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of COLING '04*.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Hirano, T., Matsuo, Y., and Kikui, G. (2007). Detecting semantic relations between named entities in text using contextual features. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 157–160. Association for Computational Linguistics.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

Joshi, M. and Penstein-Rosé, C. (2009). Generalizing dependency features for opinion mining. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 313–316.

Kywe, S. M., Hoang, T.-A., Lim, E.-P., and Zhu, F. (2012). On recommending hashtags in twitter networks. In *Proceedings of the 4th International Conference on Social Informatics*, SocInfo'12, pages 337–350, Berlin, Heidelberg. Springer-Verlag.

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM J. Res. Dev.*, 1(4):309–317.

Mazzia, A. and Juett, J. (2011). Suggesting hashtags on twitter. In *EECS 545 Project, Winter Term, 2011. URL http://www-personal.umich.edu/ amazzia/pubs/545-final.pdf*.

Ozdikis, O., Senkul, P., and Oguztuzun, H. (2012). Semantic expansion of hashtags for enhanced event detection in twitter. In *Proceedings of the 1st International Workshop on Online Social Systems*.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Salton, G. and McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., and Demirbas, M. (2010a). 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.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., and Demirbas, M. (2010b). 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.

Tisserant, G., Roche, M., and Prince, V. (2013). Gendesc : Vers une nouvelle reprsentation des donnes textuelles. *RNTI*.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.