Detecting Factual and Non-Factual Content in News Articles

Ishan Sahu
Indian Statistical Institute Kolkata
203 B T Road, Kolkata 700 108, India
ishan.sahu@gmail.com

Debapriyo Majumdar
Indian Statistical Institute Kolkata
203 B T Road, Kolkata 700 108, India
debapriyo@isical.ac.in

ABSTRACT

News articles are a major source of facts about the current state and events of our surrounding world. However, not all news articles are equally rich in presenting the facts. In this paper, we consider the problem of detecting factual and non-factual parts in news articles. We present a comprehensive survey on the existing literature on fact classification on news articles as well as a related and more widely studied problem of subjectivity vs objectivity classification of statements. Combining these techniques and some new features we design a framework for classifying facts and non-facts in news articles. We present extensive experiments on this task using several features and combinations of those on two datasets, one of which was used for subjectivity classification in previous works. We show that standard textual dataset dependent features such as n-grams produce good results on both datasets, but more general features such as part of speech tags and entity types produce inconsistent results. We analyze the results based on the nature of the datasets to present insights on the usefulness of the features and their applicability in the classification task we are considering.

1. INTRODUCTION

News articles are a major source of information about the current state and events of our surrounding world. One of the main goals, though not the only goal is to present facts. This work investigates the various features that may lead to better detection of factual content in the news articles.

1.1 Motivation

The ease and level of reach provided by the internet has resulted in massive growth of news articles. Apart from public and commercial organization based sources, smaller and even individual based sources have come up. The distinction between professional and amateur journalism has blurred. This has provided many new challenges as well as opportunities especially in the domain of analysis of the content of these articles.

An improvement in the their processing and analysis can have wide ranging impact and applications. Consider search and ranking, having a better idea about the content will definitely enhance the search results. We can also design systems that can automatically determine credibility of the article if we could identify and process the factual claims made by it. It will also help in summarization of facts associated with an event from all different sources and in presentation of a concise overview. Applications requiring information extraction from the articles can be improved. To achieve these goals we need to devise better ways to analyse the content.

News Article Snippet: The hydrocarbon sector performed poorly with crude oil (-3.3%) and natural gas (-6.9%) reporting decline in output and refineries posting a tepid 1.2% rise in May after a 17.9% rise in April. “Overall industrial activity is relatively weak. A cause for concern is weak capex, construction and realty. Road projects are picking up, which might offset in a few months.”

Factual Content	**Non-Factual Content**
1. The hydrocarbon sector performed poorly with crude oil (-3.3%) and natural gas (-6.9%) reporting decline in output and refineries posting a tepid 1.2% rise in May after a 17.9% rise in April.	1. Overall industrial activity is relatively weak. 2. A cause for concern is weak capex, construction and realty. 3. Road projects are picking up, which might offset in a few months.

Table 1: Example of factual and non-factual content from a news snippet.

A news article may contain various types of contents, such as, facts, inferences from facts, interpretations, opinions, predictions, beliefs, etc. Each of these categories has specific characteristics. Depending upon the category of an element and the intended applications, the further processes can be

CODS '17, March 09-11, 2017, Chennai, India
© 2017 ACM. ISBN 978-1-4503-4846-1/17/03…$15.00
DOI: http://dx.doi.org/10.1145/3041823.3041837

designed. Identifying and categorizing elements of an article can be considered as one of the basic steps of many different kinds of content analyses. The step seems intuitive and reasonable. For many purposes, as discussed before, designing a sieve that can separate factual content from the rest of non-factual content (including inferences, interpretations, opinions, etc.) will be a good starting point. Our goal in this paper is to detect the factual and non-factual content in news articles.

For developing algorithms to detect facts, we need to first define facts. The notion of fact is widely known and there are several definitions. For our work, we considered the following definition which is simple and easy to perceive: *A fact is something that has occurred or is actually correct.* In the context of news articles, events that have actually happened and statements which *claim* to be true are factual in nature, whereas opinions and interpretations are not.

For example, let us consider the snippet[1] in the table 1. We can observe that it contains both factual statements and non-factual content like inferences, interpretations, and predictions.

It is important to note, however, that determining whether the stated events or statements are actually true or not (hence, whether those are actually facts or not) are outside the scope of this work.

1.2 Our contribution

In this paper, we study the present state of the art on fact detection as applied specifically to news articles. We use a combination of features used in previous works and some new features to present a method for classifying fact vs non-fact in news articles. We experiment with several combination of these features on two datasets. We show that some of the features produce good results on both datasets and some of the features produce inconsistent results. We analyze the results to present insights on the usefulness of the features and their applicability in the classification task we are considering.

2. RELATED WORK

In this section, we discuss prior works which are related to fact detection on news data.

2.1 Subjectivity and Objectivity Classification

Subjectivity and objectivity studies are closely related to this work. However, our goals have subtle but important differences. We have a greater focus on the identification of factual content and its separation from the remaining part. Objectivity often also includes unprejudiced reporting and fairness in writing.

One of the earlier works on developing sentence level subjective and objective classifiers was published by Wiebe and Riloff [30]. As they worked with unannotated data, they first implemented rule-based classifiers based on general subjectivity clues to generate training data for subsequent learning algorithms. Considering patterns designed for information extraction to be effective representation of subjectivity expressions, such patterns were learned over this training data. This was based on an earlier work by Riloff and Wiebe [18]. Then, Naive Bayes classifiers were trained using these patterns and other features including subjective clues and POS features. A variant of self-training was used to improve the sentence classifier.

The classifier obtained after retraining on the new training set was reported to have subjective precision and recall as 71.3% and 86.3% respectively. Objective precision and recall were 77.5% and 57.5% respectively. Though the performance of the subjective part seems impressive, improvements in objectivity detection and recall seems imperative.

Chenlo and Losada [3] have carried out a detailed empirical study of features for sentence level subjectivity classification and polarity classification. They experimented and jointly evaluated features that have otherwise been tested independently. The subjectivity classification part is relevant to our work.

They studied various features including vocabulary features, positional features, part-of-speech features, syntactic patterns, sentiment lexicon features, features based on rhetorical structure theory (RST), concept-level features, and length features. Experiments were done on different datasets of product reviews and news articles separately. They concluded that unigrams/bigrams combined with sentiment lexicon features consistently give good performance for subjectivity classification. Once these are included, the effect of any other feature is negligible.

Biyani et al. [2] studied methods to predict subjectivity orientation of online forum threads for use in improving their retrieval. They showed that the task of identification of subjective and non-subjective discussion threads can approached using simple features generated from n-grams and part-of-speech tags. Sagae et al. [20] presented a data-driven approach to identify subjective passages that express mental and emotional states of the narrator in personal narratives. Jayawardene [9] explored paragraph level subjectivity and objectivity content analysis of online news reporting on American health care reforms.

2.2 Fact Processing and Classification

An approach to classify complete news articles into categories of either fact or opinion was put forward by Stepinski and Mittal [23]. Each sentence was classified as factual or opinion using Passive Aggressive algorithm trained on unigram, bigram, and trigram features. The overall score for an article was computed on the basis of these labels.

The dataset was obtained by crawling online news sources. Aricles with URLs containing "opinion", "editorial", and "oped" were considered as opinion-based articles. Those with "science","business", "world" were considered fact-based articles. An oversimplified assumption was made that each sentence belongs to the same class as assigned to the article containing it. News articles are usually a mix of both facts and opinions in varying degrees. Iterative training by modifying the training set and retraining with a larger set of features was used to come up with a classifier. The average F1 score of the 5-fold cross validation on the iterated training set was reported as 85%. Using additional features like POS labels and article length reduced the F1 score to 80%. As their aim was to classify complete articles, their classifier could take advantage of the fact that "sentence classification mistakes can still be overlooked as long as significant part of the article is labelled."

Hassan, et al. [7, 8] developed fact classifiers with an aim to automate fact checking in political scenarios and debates.

[1] Source: http://economictimes.indiatimes.com/articleshow/52999075.cms (accessed on July 1, 2016)

They tried to classify sentences in US presidential debates into three categories: Non-Factual Sentences (NFS), Unimportant Factual Sentences (UFS) and Check-worthy Factual Sentences (CFS). The check-worthy factual sentences were to directed to fact checkers. 79% precision and 74% recall on the CFS class was reported on the manually labelled dataset of sentences spoken by presidential candidates in past general election presidential debates. They studied features like sentiment, length, word, part-of-speech tags, and entity types.

The work by Oraby, Reed et al. [13] explored distinguishing factual arguments from emotional arguments in online dialogue. Using an annotated set of factual and feeling debate forum posts, patterns that are highly correlated with factual and emotional arguments were extracted. Then a bootstrapping method was used to extract more new patterns from unannotated posts. A post was labelled as Factual or Feeling if it matched at least three high-precision patterns for that category. They observed that factual arguments often include topic-specific terminologies, explanatory languages, and argument phrases. In contrast, the patterns related to feeling based arguments are often based on speaker's own beliefs or claims, involve assessment of the arguments by other speakers. Such arguments are also very creative and diverse.

Regmi and Bal [16] proposed a framework to determine facts and opinions in news media. Their approach was dependent on a lexicon of factual verbs and expressions which were used to distinguish between facts and opinions. The work was limited in a number of ways. Their dataset was a small corpus from editorials with low number of facts. They reported higher misclassification cases for facts.

Kastner and Monz [11] tackled the problem of extracting only the most important facts from a news article to automatically generate news highlights. Their approach was based on keyword extraction and summarization. Since they were interested only in main points of the article to produce highlights, proper identification and distinction of facts from non-facts was not required and not therefore not properly dealt with.

2.3 Other Related Works

Balahur, Steinberger et al. [1] published an approach for mining opinions from quotations in newspapers. Their work involved only direct reported speech, that is, quotations with the assumption that "quotes are usually more subjective than the other parts of the news articles". The aim was to categorize quotations for subjectivity (neutral vs. subjective) and to determine the polarity of the subjective quotations. The classification was based on subjectivity indicators. Various resources like WordNet Affect, SentiWordNet, MicroWNOp, etc. were used for polarity determination.

Wiebe and Mihalcea [29] studied subjectivity as a property of word senses and experimented on ways to improve the accuracy of word sense disambiguation for the words that have both subjective and objective senses. Su and Markert [24] also explored automatic detection of the subjectivity of word senses. Riloff et al. [19] demonstrated the use of subjectivity classification to improve the precision of information extraction.

Turney [26] published a simple unsupervised learning algorithm for classifying reviews as either recommended or not recommended. He proposed patterns of part-of-speech tags to extract subjective phrases that can be useful for determining semantic orientation. Pang and Lee [14] have done extensive work in opinion mining and sentiment analysis of online reviews and personal blogs. Tsytsarau and Palpanas [25] have also done a detailed survey on mining subjective data.

A framework was suggested by Park and Cardie [15] for automatically classifying propositions in an argument as unverifiable, verifiable non-experiential, or verifiable experiential, where the appropriate type of support is reason, evidence, and optional evidence, respectively.

Saurí and Pustejovsky [21] studied event factuality in natural language. Event factuality or the level of information expressing the factual nature of the eventualities mentioned in the text, was measured along two parameters: the notions of degree certainty (example, possible, probable, etc.) and polarity (positive, negative) of the events. Chantal van Son et al. [28] examined extraction and interpretation of perspectives on events using sentiment and event factuality. They suggested to combine the dimensions of factuality and opinion for the purpose. To deal with the problem of tracking and reconstructing news on topics spanning over a long period, Marieke van Erp et al. [27] presented a framework to model stories from news and visualisation of these storylines.

2.4 Discussion

We make few observations on the level of applicability of existing works to our aim of identifying facts in news articles and separating them from the rest of the content.

- Most of the related works focus more on the subjective content as they aim at sentiment and opinion mining.
- News articles have certain characteristics which distinguishes it from other texts. There is no comprehensive study targeted on detecting fact and non-fact from news articles.
- Stepinski and Mittal [23] classified complete news articles. For better content analyses, we need a finer level processing. It is important to note that mainstream news articles are seldom either completely factual or completely non-factual. A granular fact classification method for parts of an article would also enable comparison of different news sources in terms of the amount of fact presented in their articles.

2.5 Difference between factual and objective content

Though there is clearly a relationship between factual and objective content, there are subtle differences which may be important in our case. Objectivity in reporting includes elements of fairness and absence of bias. Oraby et al. [13] observes "There is clearly a relationship between a proposition being FACTUAL versus OBJECTIVE or VERIDICAL, although each of these different labelling tasks may elicit differences from annotators."

3. IDENTIFYING FACT VS. NON-FACT

Based on previous works and our general observation, we believed that certain features will show differences across categories of content. We intended to study and utilise them for our fact identification and classification. A framework was designed so that we can test the various combinations of features for the purpose and observe their efficacy.

The process followed during both training of the classifier and classifying content using the trained classifier is similar. It follows the following steps: (1) unit segmentation, (2) feature extraction, and (3) classification.

3.1 Unit segmentation

The whole article was broken down into smaller and simpler units for processing. The segmentation was done on two levels:

1. Sentences
2. Direct Quotations

Direct quotations here are defined as phrases/sentences enclosed within various forms of quotation marks. We considered quotations of first level depths only.

First, the sentence boundaries and the positions of direct quotations were identified. Then if quotations were encountered in a sentence, the quotations were separated from the remaining part of the sentence and segmented as separate units. Also, multiple sentences in a direct quotation were segmented as separate units. For example, table 2 presents a news article snippet and then shows its corresponding segmentation into units as defined here.

News Article Snippet: Officials said two civilians were also injured in the attack. “We are heartbroken,” Brown said during a news conference Friday. “There are no words to describe the atrocity that occurred to our city.” [2]

Segmented Units

1. Officials said two civilians were also injured in the attack.
2. “We are heartbroken,”
3. Brown said during a news conference Friday.
4. “There are no words to describe the atrocity that occurred to our city.”

Table 2: Partitioning news article into units for processing

We used annotators in Stanford CoreNLP [12] to identify sentences and first level quotations. Using this information, we segmented the text into units for further processing.

3.2 Feature extraction

The feature extraction step includes extraction of several features, namely N-grams, Part-of-Speech (POS) Tags, entity types, AutoSlog patterns, subjective patterns, sentiment, positional features and POS-patterns.

3.2.1 N-grams extraction

Word level n-grams capture the use of language and may be useful in identifying word sequences more prevalent in a specific category of content. For example, sequences containing words such as “believe”, “hope”, etc. are most likely to be non-factual whereas those containing “reported”, “occurred”, etc. are more likely to be factual. Other works [23, 7, 3] have studied the use of n-grams in similar classification tasks. We built lists of word level unigrams, bigrams, and trigrams processed over the training data. These sequences of words were used as dimensions for the n-grams feature group. The frequency of an n-gram was taken as the value of the dimension corresponding to that n-gram for the respective unit.

3.2.2 Part-of-speech (POS) tags extraction

In human languages, many words have different senses based on their usage in different parts-of-speech. For example, “content” as a verb has a single sense related to satisfaction, whereas as a noun apart from the sense related to satisfaction, it can refer to things that are contained in something. So, when used as verb, it has higher chances of characterizing non-facts. Whereas, if as a noun used to describe the composition of some thing, it would be denoting facts. Various studies [30, 3, 7, 23], have worked with POS tags to solve similar problems.

Tag	**Meaning**
JJ	Adjective
NN	Noun, singular or mass
WRB	Wh-adverb
PRP	Personal pronoun
DT	Determiner
NNP	Proper noun, singular
FW	Foreign word
NNS	Noun, plural

Table 3: A subset of Part-of-Speech Tag features

A list of part-of-speech tags present in the training data was generated and those tags were used as this feature group’s dimensions. At the time of feature extraction of a given unit, the count of each POS tag in the unit was assigned to the corresponding dimension as its value. Stanford CoreNLP [12] was used for POS tagging. See table 3 for a subset of the POS tags that were considered.

3.2.3 Entity types extraction

Presence of entities of specific type may indicate the category of text we are dealing with. Intuitively, time, date, money, number, etc. type of entities will be more common in facts and can help differentiate it from non-facts in news articles. For example, if money and organization entities are present together in a unit, it may be stating something about the finances of the organization. Similarly, if time and date are present in a sentence, the probability of the sentence presenting some incident is higher than otherwise. Hassan

[2] *Source: https://www.washingtonpost.com/news/morning-mix/wp/2016/07/08/like-a-little-war-snipers-shoot-11-police-officers-during-dallas-protest-march-killing-five/ (accessed on July 8, 2016)*

et al. [7] used entity type features for fact classification in US presidential debates.

Similar to the approach with POS tags, the entity types encountered in the training data were considered as dimensions for this feature group. During feature extraction, the count of each entity type was generated.

Entity type annotation was performed using Stanford CoreNLP [12] using the default models. In our studies, all the entity types detected by the tool was present in the data and so all types were included. See table 4 for the entity types considered.

	Entity Types
1	LOCATION
2	NUMBER
3	MONEY
4	PERSON
5	SET
6	MISC
7	TIME
8	ORDINAL
9	ORGANIZATION
10	DATE
11	PERCENT
12	DURATION

Table 4: Entity types

3.2.4 AutoSlot-TS patterns (ASPattern) extraction

Patterns that help in extracting targeted types of noun phrases are useful in information extraction tasks. They are used to get the factual data of specific types from large texts. If we can learn every extraction pattern for noun phrases relevant to factual content, we can use them to characterize factual content. Also, patterns used to extract information from texts have been shown to represent expressions associated with subjectivity [19, 30]. Oraby et al. [13] also proposed an approach to classify arguments in online dialogue using such patterns.

Information extraction patterns relevant to fact class were generated with associated statistics using AutoSlog-TS [17]. We will refer the set of such patterns in our analysis as "ASPattern". Out of all the patterns generated, we filtered the ones with minimum frequency value, $\theta_f = 3$, and minimum probability value, $\theta_p = 0.70$, similar to the parameter values followed by [13]. The filtered patterns formed the dimensions for this feature group.

The number of times an IE pattern was present in the unit, was taken as the value of the corresponding dimension for the respective unit.

3.2.5 Subjective patterns (Tpattern) extraction

Simple part-of-speech patterns denoting some kind of opinion or subjectivity can help characterize the non-factual part of the news. For example, patterns involving adjectives are more likely to be opinions. The phrase "a breathtaking journey" which matches some adjective patterns, expresses a feeling. Turney [26] used such patterns to identify phrases for detecting semantic orientation. Chenlo et al. [3] have used those patterns to study similar classification.

All the five Turney's patterns [26] were included in this feature group and are referred together as "Tpattern". See table 5 for the evaluative phrase identifying patterns proposed by Turney. The number of times a pattern was present in the unit was set as the value of the dimension corresponding to the pattern.

Table 5: Patterns of tags for extracting phrases with possible semantic orientation

	First Word	**Second Word**	**Third Word (Not Extracted)**
1	JJ	NN or NNS	anything
2	RB, RBR, or RBS	JJ	not NN nor NNS
3	JJ	JJ	not NN nor NNS
4	NN or NNS	JJ	not NN nor NNS
5	RB, RBR, or RBS	VB, VBD, VBN, or VBG	anything

3.2.6 Sentiment extraction

Intuitively, non-facts including opinions in news articles exhibit positive or negative sentiments whereas facts show no sentiments or are neutral. Chenlo et al. [3] used lexicon based sentiment features for their classification. A score depending upon the sentiment of the sentence was used by Hassan et al. [7].

We used models by Socher et al. [22] as implemented in Stanford CoreNLP to get sentiment information at the unit level. The models are basically a type of Recursive Neural Network that builds on top of grammatical structures. Each point on the 5-point scale of 0 = very negative, 1 = negative, 2 = neutral, 3 = positive, and 4 = very positive, was considered as a dimension. The scores for each of these dimensions were obtained using the sentiment models.

3.2.7 Positional features extraction

News articles generally has facts followed by analysis and comments. Comments are generally non-factual which suggests that non-facts are more concentrated towards the later parts of the article. Kastner and Monz [11] utilized position of the sentence in the text to extract most important facts. Chenlo et al. [3] also studied positional features in their subjectivity and polarity classification. It will be interesting to see its applicability in our case.

Each unit has a location in the article. We defined six types (three absolute positions and their three normalized versions) of positional features:

1. Unit Number: The number of the unit from the beginning of the news article.
2. Normalized Unit Number: Unit Number normalized over the total number of units.
3. Starting Character Position: The position of the starting character of the unit from the top of the article.
4. Normalized Starting Character Position: Starting character position normalized over the total number of characters in the article.

5. Ending Character Position: The position of the ending character of the unit from the top of the article.

6. Normalized Ending Character Position: Ending character position normalized over the total number of characters in the article.

3.2.8 *POS patterns*

The topic for different news articles could be different, but the formulation of sentences to present facts could retain some similarity across different news articles. To capture such patterns, we used the part of speech (POS) patterns. We replaced the individual words in every sentence by their parts of speech (POS) to obtain the POS-sequences and constructed N-gram patterns of length 3, 4 and 5 from the POS sequences. These N-grams are used as features in our classification. The choice of a suitable length of N-grams is determined experimentally and we found that lengths 3 and 4 produce relatively better results.

3.3 Classification

While training, the generated vectors on training data was used to train the classifier. On classification, the vectors were fed to the classifier which then produced the predictions.

We used Support Vector Machine (SVM) for classification. SVMs perform well for text classifications [10]. Text data usually belong in high dimensional space and have few irrelevant features. Their vector representations are sparse. Also, most text categorization problems are linearly separable.

The LIBLINEAR package [6] is a simple package for solving large-scale regularized linear classification and regression. We used the L2-regularized L2-loss support vector classification in primal formulation.

4. EXPERIMENTAL RESULTS

We carried out experiments on classifying facts vs nonfacts using different combinations of the features described in Section 3. The datasets used along with the experiment setup and their results are presented in this section.

4.1 Datasets

We performed our experiments on two datasets, namely the MPQA Opinion Corpus[3] [31, 5] and the Signal Media One-Million News Articles Dataset[4] [4].

4.1.1 *The MPQA Opinion Corpus*

The MPQA Opinion Corpus contains news articles and other text documents manually annotated for opinions and other private states (i.e., beliefs, emotions, sentiments, speculations, etc.). MPQA 3.0 consists of 70 documents, a subset of previous MPQA. We used the 16 annotated files in GATE to prepare our dataset. The types of annotations that are provided in the documents of this corpus as explained in their documentation are:

1. `agent` annotation: Marks phrases that refer to sources of private states and speech events, or phrases that refer to agents who are targets of an attitude.

2. `expressive-subjectivity` annotation: Marks expressive-subjective elements, words and phrases that indirectly express a private state.

3. `direct-subjective` annotation: Marks direct mentions of private states and speech events (spoken or written) expressing private states.

4. `objective-speech-event` annotation: Marks speech events that do not express private states.

5. `attitude` annotation: Marks the attitudes that compose the expressed private states.

6. `targetFrame`: Records the span-based target annotations and entity/event-level annotations for each attitude, expressive subjectivity and objective speech event. Automatically generated.

7. `sTarget` annotation: Marks the span-based targets of the attitudes, i.e., what the attitudes are about or what the attitudes are directed toward. Previously the annotations are named as “target”.

8. `eTarget` annotation: Marks the entity/event-level target of the attitudes, expressive subjectivites and objective speech events. The eTarget is anchored to a noun phrase head or a verb phrase head.

9. `sentence` annotation: Marks each sentence.

10. `supplementaryAttitude` annotation: Marks the attitudes that compose the expressed private states, that were identified when developing MPQA 3.0 version.

11. `supplementaryExpressive-subjectivity` annotation: Marks expressive-subjective elements, words and phrases that indirectly express a private state, that were identified when developing MPQA 3.0 corpus.

Works on subjectivity classification [30, 3] have used MPQA corpus for evaluations. Since it does not have annotations defined for facts, we adapted it to our settings. The scheme we followed is described below:

1. If the unit contained only annotations of type `direct-subjective`, `expressive-subjectivity`, and `attitude`, it was labelled as non-fact.

2. If the unit contained only `objective-speech-event` annotation, it was labelled as fact.

3. If the unit contained both types of annotations mentioned in above two points, the conflict was manually resolved.

4. The units not containing any of the above types of annotations was labelled as fact. This was a coarse simplification based on the belief that subjectivity clues will be absent in factual units. This step was taken primarily to reduce the imbalance between the number of fact and non-fact units.

We were able to generate a total of 557 units out of which 167 were labelled as facts and 390 as non-facts.

[3] http://mpqa.cs.pitt.edu/corpora/mpqa_corpus/

[4] http://research.signalmedia.co/newsir16/signal-dataset.html

4.1.2 *The Signal Media One-Million News Articles Dataset*

The Signal Media One-Million News Articles Dataset is released by Signal Media to facilitate conducting research on news articles. The articles of the dataset were collected from a variety of news sources for a period of 1 month (1-30 September 2015). It contains 1 million articles that are mainly English, but they also include non-English and multi-lingual articles. Sources of these articles include major ones, such as Reuters, in addition to local news sources and blogs. Each article has the following fields:

1. `id`: a unique identifier for the article
2. `title`: the title of the article
3. `content`: the textual content of the article (may occasionally contain HTML and JavaScript content)
4. `source`: the name of the article source (e.g. Reuters)
5. `published`: the publication date of the article
6. `media-type`: either "`News`" or "`Blog`"

The number of individual unique sources are over 93k. The dataset contains 265,512 Blog articles and 734,488 News articles. The average length of an article is 405 words.

As the MPQA dataset had certain limitations based on the scheme it was generated, we also experimented on labelled dataset prepared using the Signal Media dataset. Several articles of media-type "`News`" were taken. The sentences from these articles were given to 8 annotators, 6 of them returned labelled data. The annotators were well-educated people but not computer scientists and unaware of text classification algorithms, in general. They were asked to label each unit by factual, non-factual or leave it blank if they are not sure. They were given the instruction to label the sentences based on the definition in Section 1. From the consolidated inputs, only units for which a label has a vote of at least 3-1 margin were taken. Overall, there were 377 labelled units (203 factual, 174 non factual).

4.2 Experimental setup

Each dataset was divided into two parts: training set containing 75% of units and test set containing 25% of units. We experimented with the linear support vector classifier of LIBLINEAR package (L2-regularized L2-loss support vector classification (primal)) [6]. The evaluation process is as follows:

1. Each of the features of the training data were scaled to be in $[0, 1]$ and the scaling factors were obtained. Using the same scaling factors, test data was scaled.
2. Parameter search was done with 10-fold cross validation on the training data. Optimum value of cost parameter was learned.
3. Using the cost value obtained above, linear model was trained on the training data.
4. The model was then used to generate predictions for the test set.
5. Various statistics were computed by comparing predictions with the actual labels.

The process was used with various combinations of feature groups and results were obtained.

4.3 Evaluation Metrics

Let, U be the set of units that was included in the classification task, and T_f be the set of facts, and T_{nf} be the set of non-facts, such that, $T_f \cup T_{nf} = U$ and $T_f \cap T_{nf} = \phi$. Let P_f be the set of units that the classifier predicted to be facts, and P_{nf} be the set of units predicted to be non-facts. Then, our metrics for evaluataion are:

1. Fact classification metrics:

$$\text{Fact Precision, } p_{fact} = \frac{|T_f \cap P_f|}{|P_f|} \tag{1}$$

$$\text{Fact Recall, } r_{fact} = \frac{|T_f \cap P_f|}{|T_f|} \tag{2}$$

$$\text{Fact } F_1 = \frac{2 \times p_{fact} \times r_{fact}}{p_{fact} + r_{fact}} \tag{3}$$

2. Non-Fact classification metrics:

$$\text{Non-Fact Precision, } p_{non\text{-}fact} = \frac{|T_{nf} \cap P_{nf}|}{|P_{nf}|} \tag{4}$$

$$\text{Non-Fact Recall, } r_{non\text{-}fact} = \frac{|T_{nf} \cap P_{nf}|}{|T_{nf}|} \tag{5}$$

$$\text{Non-Fact } F_1 = \frac{2 \times p_{non\text{-}fact} \times r_{non\text{-}fact}}{p_{non\text{-}fact} + r_{non\text{-}fact}} \tag{6}$$

3. Total accuracy:

$$\text{Total accuracy, } A = \frac{|T_f \cap P_f| + |T_{nf} \cap P_{nf}|}{|U|} \tag{7}$$

4.4 Results

In this section we present the results of our experiments. For each of the given feature combinations, we compute the precision, recall, and F_1 measure for facts and non-facts separately. The combined accuracy of classification is also measured. We also visualise the results using precision-recall graphs. Each point in these graphs correspond to a feature combination that was used in classification. The points are plotted with precision on y-axis and recall on x-axis.

The results for the dataset from MPQA is summarized in table 6. Figure 1 and figure 2 shows the precision-recall graph for fact classification and non-fact classification on the MPQA dataset, respectively. In Table 7, we show the experimental results on Signal Media dataset. Figure 3 and figure 4 shows the precision-recall graph for fact classification and non-fact classification on the Signal Media dataset, respectively.

4.5 Discussion

4.5.1 *The MPQA Opinion Corpus*

From the results stated in the previous section we can observe the following for individual features:

- N-grams by itself present a somewhat decent classification power with 50% precision and 44% recall values for facts, and 77% precision and 81% recall for non-facts.

Features	Facts			Non-Facts			
	p	r	F_1	p	r	F_1	A
Ngrams	0.50	0.44	0.47	0.77	0.81	0.79	0.70
POS	1.00	0.02	0.05	0.71	1.00	0.83	0.71
Entity	0.30	0.15	0.20	0.70	0.86	0.77	0.64
ASPattern	1.00	0.10	0.18	0.72	1.00	0.84	0.73
Tpattern	0.00	0.00	–	0.70	0.99	0.82	0.70
Sentiment	NaN	0.00	–	0.70	1.00	0.83	0.70
Position	0.50	0.02	0.05	0.71	0.99	0.82	0.70
POSPattern(4)	0.63	0.13	0.21	0.54	0.93	0.68	0.55
Ngrams + POS	0.62	0.39	0.48	0.78	0.90	0.83	0.75
Ngrams + Entity	0.50	0.41	0.45	0.77	0.82	0.80	0.70
Ngrams + ASPattern	0.50	0.44	0.47	0.77	0.81	0.79	0.70
Ngrams + Tpattern	0.53	0.44	0.48	0.78	0.84	0.81	0.72
Ngrams + Sentiment	0.71	0.24	0.36	0.75	0.96	0.84	0.75
Ngrams + Position	0.60	0.29	0.39	0.75	0.92	0.83	0.73
Ngrams + POSPattern(4)	0.71	0.26	0.38	0.57	0.91	0.70	0.60
POS + ASPattern + Tpattern	1.00	0.02	0.05	0.71	1.00	0.83	0.71
All features	0.59	0.39	0.47	0.77	0.89	0.83	0.74

Table 6: Fact vs. non-fact classification on the dataset prepared using the MPQA data. The precision (p), recall (r), the F_1 score for the facts and non-facts separately and the total accuracy (A) are shown for various combinations of features.

- Turney's patterns (Tpattern) and Sentiment features, each by themselves are unable to distinguish factual content from the non-factual content. Tpattern features alone lead to zero correct classification of facts. Using Sentiment alone, the classifier did not predict even a single unit as fact.

- AutoSlog-TS Patterns (ASPattern) has some differentiating power and can help identify facts with high precision. But it fails to recognise a lot of them and labels them as non-facts. Similar is the case with POS tags.

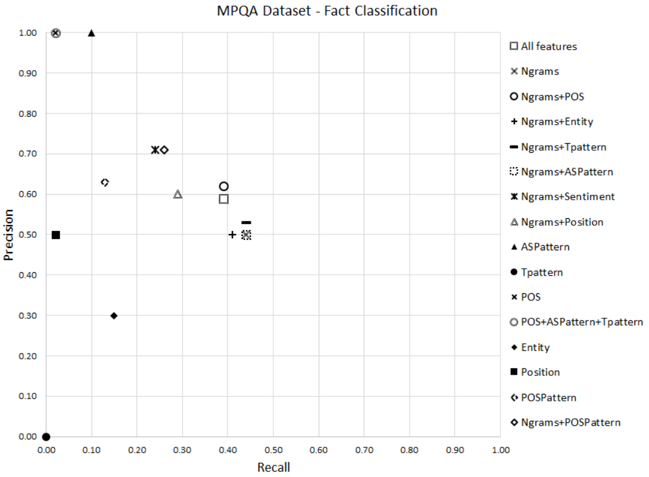

Figure 1: Precision-Recall Graph for facts in the MPQA Dataset. Each point in the graph correspond to a feature combination that was used in classification. The points are plotted with precision on y-axis and recall on x-axis. Since precision for Sentiment is not defined, it is not plotted.

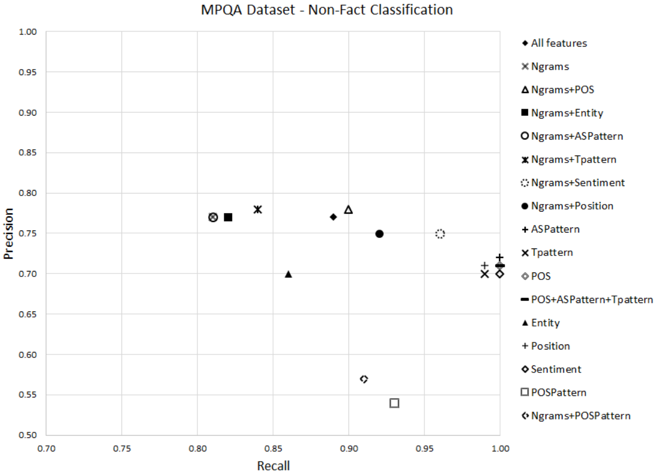

Figure 2: Precision-Recall Graph for non-facts in the MPQA Dataset. Each point in the graph correspond to a feature combination that was used in classification. The points are plotted with precision on y-axis and recall on x-axis.

- Position and Entity type features by themselves does not seem to be much useful because of their poor performance in both precision and recall.

- The POS-patterns with length 4 performed better than those with other lengths for this dataset. We reported the figures for 4-gram POS-patterns in Table 6. The POS-patterns together with sentiment are able to achieve very high precision in classifying both facts and non-facts, but suffers from poor recall for facts.

- Overall, we can conclude that feature combinations excluding n-grams doesn't seem useful in this dataset.

Features	**Facts**			**Non-Facts**			
	p	r	F_1	p	r	F_1	A
Ngrams	0.63	0.86	0.73	0.72	0.42	0.53	0.66
POS	0.76	0.70	0.73	0.68	0.74	0.71	0.72
Entity	0.75	0.72	0.73	0.69	0.72	0.70	0.72
ASPattern	0.90	0.36	0.51	0.56	0.95	0.71	0.63
Tpattern	0.45	0.42	0.43	0.37	0.40	0.38	0.41
Sentiment	0.54	0.94	0.69	0.50	0.07	0.12	0.54
Position	0.54	0.74	0.63	0.48	0.28	0.35	0.53
POSPattern(3)	0.66	0.88	0.75	0.77	0.47	0.58	0.69
Ngrams + POS	0.65	0.90	0.76	0.79	0.44	0.57	0.69
Ngrams + Entity	0.65	0.90	0.76	0.79	0.44	0.57	0.69
Ngrams + ASPattern	0.65	0.88	0.75	0.76	0.44	0.56	0.68
Ngrams + Tpattern	0.68	0.86	0.76	0.77	0.53	0.63	0.71
Ngrams + Sentiment	0.67	0.92	0.77	0.83	0.47	0.60	0.71
Ngrams + Position	0.67	0.88	0.76	0.78	0.49	0.60	0.70
Ngrams + POSPattern(3)	0.66	0.84	0.74	0.72	0.49	0.58	0.68
POS + ASPattern	0.77	0.66	0.71	0.66	0.77	0.71	0.71
POS + ASPattern + Tpattern	0.76	0.68	0.72	0.67	0.74	0.70	0.71
POS + Sentiment	0.74	0.74	0.74	0.70	0.70	0.70	0.72
All features	0.69	0.86	0.77	0.77	0.56	0.65	0.72

Table 7: Fact vs. non-fact classification on Signal Media dataset. The precision (p), recall (r), the F_1 score for the facts and non-facts separately and the total accuracy (A) are shown for various combinations of features.

Experimenting using additional features with N-grams, we see small improvements in some cases, while slight deterioration in others.

- We do not witness any improvement in recall values for facts.

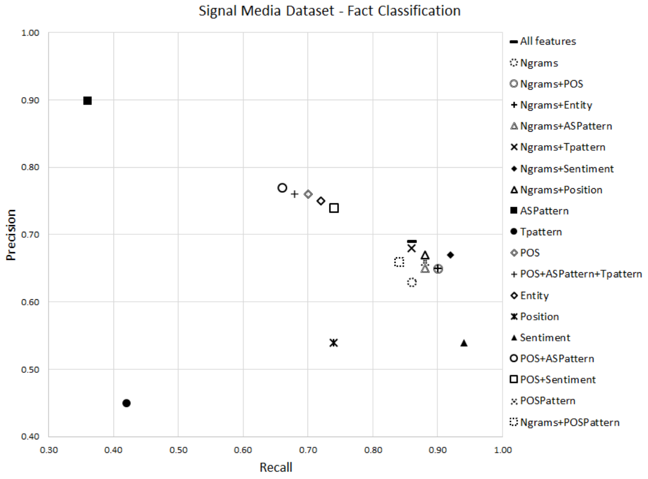

Figure 3: Precision-Recall Graph for facts in the Signal Media Dataset. Each point in the graph correspond to a feature combination that was used in classification. The points are plotted with precision on y-axis and recall on x-axis.

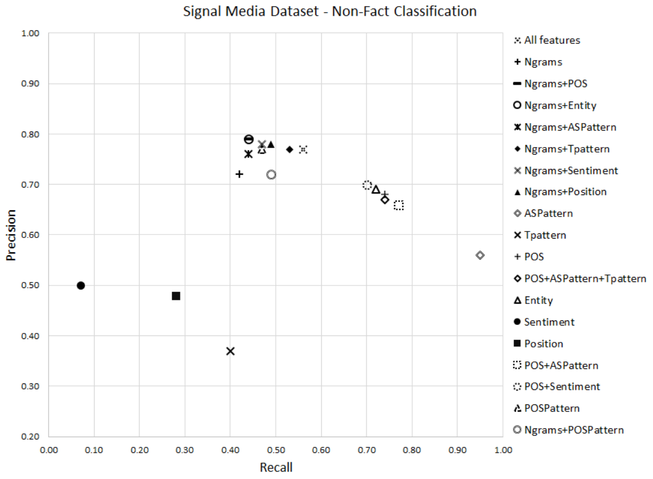

Figure 4: Precision-Recall Graph for non-facts in the Signal Media Dataset. Each point in the graph correspond to a feature combination that was used in classification. The points are plotted with precision on y-axis and recall on x-axis.

- Using POS tags' count there is an 12% increase in precision but about 5% decrease in recall.
- Including Entity types leads to a minor decrease in fact recall performance by 3% approximately.
- With Tpattern, there is only a small improvement of about 3% in precision while recall staying the same.
- N-grams and sentiments show a high improvement in precision of labelling facts (71%). But the recall rates are considerably lowered. With positional features we see similar trend but lesser than what we see with sentiments.

- Using all the features improves the precision of fact classification by about 9% but the recall decreases by about 5%.
- The performance regarding non-facts stays decent in all these cases.

4.5.2 *The Signal Media One-Million News Articles Dataset*

For the experiments involving individual feature groups, we notice the following:

- N-grams have good performance except that the recall for non-facts is low.
- POS tags and Entity type features by themselves demonstrate good performance for both facts as well as non-facts.
- Use of ASPattern leads to high precision in facts and high recall in non-facts.
- The POS-patterns with length 3 performed better than those with length 4 and 5 for this dataset and achieved decent precision and recall.
- Tpattern and Positional features does not deliver well and does not appear to be useful alone.
- Sentiment based classification leads to highest recall (94%) for facts in all our experiments. But it suffers from low precision of 54%.

The trends observed while studying feature combinations are:

- For n-gram combinations, there is only a slight improvement in precision and recall values for both facts and non-facts. The recall for non-facts is still low. Single features like POS and Entity still delivers better performance than any of these compositions with regards to precision. However, n-gram compositions may help when high fact recall is useful.
- Trying to combine high fact precision feature POS and high fact recall Sentiment, we observe classifier with balanced metrics.

Comparison of Results in the Two Datasets

We notice that n-grams and its combinations offer decent performance for both the datasets. Also, ASPattern gives high fact precision classifiers. The performance of positional features also does not look encouraging in both the cases.

We see some striking differences in the experiment results on the two datasets. There is a marked improvement in performance of POS and Entity features. Even Sentiment results seem better.

5. CONCLUSION AND FUTURE WORK

We experimented with various features and their combinations to study their usefulness in identifying facts in news articles. The evaluation was done using two datasets: one prepared using the MPQA Opinion Corpus, and the other created by getting annotations on the Signal Media dataset.

Experimental results were similar for some features and very different for some other features on these two datasets. This may be due to the absence of proper annotation signifying factual content in the MPQA dataset and the approximate scheme used to generate factual units. The Signal Media dataset was annotated with focus on the present task of fact identification and was therefore, a better quality dataset. The much improved performance in the second case indicates that proper dataset is essential for fact identification studies.

We showed that N-grams and some of its combinations perform well in case of both datasets. However, since n-grams are generated over the part of corpus under study, it may have certain domain dependencies. A model trained on one dataset may not work well on another dataset. On the other hand, we find good performances for the more general Part-of-Speech (POS) and Entity features on the Signal Media dataset. As these features are less dependent on domain of news articles than n-grams, they can help design more generalised classifiers for facts. Cross dataset experiments involving several news article datasets would be an important future work to determine whether a general model trained on POS and entity features can be used to classify facts in case of general news articles.

AutoSlog-TS based information extraction patterns seem to provide high fact precision which might be due to the fact that the patterns are learned on the dataset. Cross dataset studies will be useful here as well. Turney's patterns does not help appreciably in fact identification.

Sentiment seem to provide some useful information in the high quality dataset and its combinations with other feature groups may be further studied. From the below par performance of positional features and its combinations, they can be safely ignored while designing fact classifiers.

Though we tried to identify facts at a finer level by breaking down sentences and quotations, it had limitations. We feel that clause level segmentation will be an interesting future work.

The experience during experiments made us realise the lack and importance of high quality fact annotated dataset of news articles. Developing more datasets for facts classification on news articles and further experiments with larger datasets is required.

The detection of facts can be extended to develop applications for assessing the correctness and credibility of news articles. A mature fact detection system would be able detect facts from news articles from different sources and it would be an interesting work to compare them.

6. REFERENCES

[1] A. Balahur, R. Steinberger, E. v. d. Goot, B. Pouliquen, and M. Kabadjov. Opinion mining on newspaper quotations. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 03*, WI-IAT '09, pages 523–526, Washington, DC, USA, 2009. IEEE Computer Society.

[2] P. Biyani, C. Caragea, A. Singh, and P. Mitra. I want what i need!: Analyzing subjectivity of online forum threads. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2495–2498, New York, NY, USA, 2012. ACM.

[3] J. M. Chenlo and D. E. Losada. An empirical study of sentence features for subjectivity and polarity

classification. *Information Sciences*, 280:275–288, 2014.

[4] D. Corney, D. Albakour, M. Martinez, and S. Moussa. What do a million news articles look like? In *Proceedings of the First International Workshop on Recent Trends in News Information Retrieval co-located with 38th European Conference on Information Retrieval (ECIR 2016), Padua, Italy, March 20, 2016.*, pages 42–47, 2016.

[5] L. Deng and J. Wiebe. Mpqa 3.0: An entity/event-level sentiment corpus. In *Conference of the North American Chapter of the Association of Computational Linguistics: Human Language Technologies*, 2015.

[6] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[7] N. Hassan, B. Adair, J. T. Hamilton, C. Li, M. Tremayne, J. Yang, and C. Yu. The quest to automate fact-checking. *world*, 2015.

[8] N. Hassan, C. Li, and M. Tremayne. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1835–1838, New York, NY, USA, 2015. ACM.

[9] W. Jayawardene. A content analysis of online news media reporting on american health care reform. In *Proceedings of the International Conference on Data Mining (DMIN)*, page 1. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012.

[10] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer, 1998.

[11] I. Kastner and C. Monz. Automatic single-document key fact extraction from newswire articles. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 415–423, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[12] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.

[13] S. Oraby, L. Reed, R. Compton, E. Riloff, M. Walker, and S. Whittaker. And that's a fact: Distinguishing factual and emotional argumentation in online dialogue. *NAACL HLT 2015*, page 116, 2015.

[14] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, Jan. 2008.

[15] J. Park and C. Cardie. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[16] S. Regmi and B. K. Bal. What make facts stand out from opinions? distinguishing facts from opinions in news media. In *Creativity in Intelligent, Technologies and Data Science*, pages 655–662. Springer, 2015.

[17] E. Riloff and W. Phillips. An introduction to the sundance and autoslog systems. Technical report, Technical Report UUCS-04-015, School of Computing, University of Utah, 2004.

[18] E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 105–112, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[19] E. Riloff, J. Wiebe, and W. Phillips. Exploiting subjectivity classification to improve information extraction. In *Proceedings of the national conference on artificial intelligence*, volume 20, page 1106. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

[20] K. Sagae, A. S. Gordon, M. Dehghani, M. Metke, J. S. Kim, S. I. Gimbel, C. Tipper, J. Kaplan, and M. H. Immordino-Yang. A data-driven approach for classification of subjectivity in personal narratives. In *OASIcs-OpenAccess Series in Informatics*, volume 32. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2013.

[21] R. Saurí and J. Pustejovsky. Are you sure that this happened? assessing the factuality degree of events in text. *Comput. Linguist.*, 38(2):261–299, June 2012.

[22] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer, 2013.

[23] A. Stepinski and V. Mittal. A fact/opinion classifier for news articles. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 807–808, New York, NY, USA, 2007. ACM.

[24] F. Su and K. Markert. From words to senses: A case study of subjectivity recognition. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 825–832, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[25] M. Tsytsarau and T. Palpanas. Survey on mining subjective data on the web. *Data Min. Knowl. Discov.*, 24(3):478–514, May 2012.

[26] P. D. Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised 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.

[27] M. van Erp, G. Satyukov, P. Vossen, and M. Nijssen. Discovering and visualising stories in news. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, Reykjavik, Iceland, May 26-31 2014.

[28] C. van Son, M. van Erp, A. Fokkens, and P. Vossen.

Hope and fear: Interpreting perspectives by integrating sentiment and event factuality. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, Reykjavik, Iceland, May 26-31 2014.

[29] J. Wiebe and R. Mihalcea. Word sense and subjectivity. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 1065–1072, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[30] J. Wiebe and E. Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'05, pages 486–497, Berlin, Heidelberg, 2005. Springer-Verlag.

[31] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2):165–210, 2005.