Proceedings of the Tenth International AAAI Conference on
Web and Social Media (ICWSM 2016)

TweetGrep: Weakly Supervised Joint Retrieval and Sentiment Analysis of Topical Tweets

Satarupa Guha^{1,a}, **Tanmoy Chakraborty**^{2,b}, **Samik Datta**^{3,c}, **Mohit Kumar**^{3,d}, **Vasudeva Varma**^{1,e}
^{1}International Institute of Information Technology, Hyderabad, India, ^{a}satarupa.guha@research.iiit.ac.in, ^{e}vv@iiit.ac.in
^{2}University of Maryland, College Park, MD 20742, ^{b}tanchak@umiacs.umd.edu
^{3}Flipkart Internet Pvt Ltd, {^{c}samik.datta,^{d}k.mohit}@flipkart.com

Abstract

An overwhelming amount of data is generated everyday on social media, encompassing a wide spectrum of topics. With almost every business decision depending on customer opinion, mining of social media data needs to be quick and easy. For a data analyst to keep up with the agility and the scale of the data, it is impossible to bank on fully supervised techniques to mine topics and their associated sentiments from social media. Motivated by this, we propose a weakly supervised approach (named, `TweetGrep`) that lets the data analyst easily define a topic by few keywords and adapt a generic sentiment classifier to the topic – by jointly modeling topics and sentiments using label regularization. Experiments with diverse datasets show that `TweetGrep` beats the state-of-the-art models for both the tasks of retrieving topical tweets and analyzing the sentiment of the tweets (average improvement of 4.97% and 6.91% respectively in terms of area under the curve). Further, we show that `TweetGrep` can also be adopted in a novel task of *hashtag disambiguation*, which significantly outperforms the baseline methods.

1 Introduction

With social media emerging as the de facto destination for their customers' views and opinions, customer-centric organisations around the world are investing on mining social media conversations to gauge public perceptions. Twitter, one of the largest amongst these platforms, with a staggering 500 Million daily tweets and 320 Million monthly active users[1], witnessed a variety of usage ranging from a platform for political reforms to an instrument for reporting earthquakes. Owing to the relatively open nature of Twitter, it had been the subject of study in majority of the published literature, and is the platform we focus herein.

We put ourselves into the shoes of a social media analyst tasked with gauging public perceptions around her organisation (and that of her competitors, perhaps). It is seldom useful to perform sentiment analysis of *all* the tweets related to the organisation; being aggregate, this form of introspection is seldom actionable. A more desirable form of introspection, we hypothesize, would be to isolate *topical* tweets (e.g. pertaining to mobile app-only move for FLIPKART, the largest Indian e-commerce portal), and then perform sentiment analysis *only* on them.

One has to surmount two major obstacles. First, the agility of today's organisations - manifested in the form of frequent Feature Releases, and frequent Promotional Campaigns, to name a few - would require such a task to be performed daily, possibly dozens of times each day. A query-based retrieval[2] is often rendered inadequate due to the lack of fixed linguistic norms and the sheer dynamism and diversity in Twitter. Moreover, intrinsic ambiguity demands that retrievals be *context sensitive* - e.g. a keyword search by `civil war` would retrieve tweets related to *CivilWar* the motion picture, as well as, the ongoing civil war in Syria. On the other extreme, *learning* to retrieve via classification, while desirable, cannot cope with the agility owing to its demanding supervision needs. Secondly, a sentiment analyser learnt from a generic corpus of tweets often misses topic-specific connotations. Domain-adaptation, without requiring additional supervision, is much desired.

In this work, we develop a methodology (we call it `TweetGrep`) that enables the analyst to specify each such topic-of-interest with just a *pair* of queries - one retrieving *all* tweets that are even remotely related (maximising the recall), and, the other retrieving only *topical*-tweets (hence, maximising the precision). Additionally, many such topics demonstrate a dominant polarity of opinion that is apparent to a domain-specialist - e.g. while tweets related to CIVILWAR (motion picture) has predominantly positive sentiment, those related to the Syrian civil war are almost always negative. In light of this, we further request the analyst to furnish *label proportion*s: the expected opinion polarity of topical (and non-topical) tweets.

In what follows, it will be demonstrated that this mode of supervision not only aids the retrieval task, but when modeled *jointly*, reinforces the sentiment analysis as well, by forcing it learn topic specific connotations - *without additional supervision*. `TweetGrep` beats the state-of-the-art models for both the tasks of retrieving topical tweets and analyzing the sentiment of the tweets with an average improvement of 4.97% and 6.91% respectively in terms of area under the curve.

Furthermore, as an additional utility of `TweetGrep` we

Copyright © 2016, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1] As of Jan 9, 2016. Source: https://goo.gl/tX5WjU

[2] E.g. Twitter Search https://goo.gl/jIX0Ku

demonstrate its competence in the novel task of *hashtag disambiguation*. A hashtag is a type of label or metadata tag used on social networks, which makes it easier for users to find messages with a specific theme or content. Many people use hashtags to identify what is going on at events, emergencies and for following breaking news. However, in recent times, hashtags have been thoroughly abused, often overloaded, and their definitions morphed with time, so much so that real information gets drowned amidst the ocean of irrelevant tweets, all using the same hashtag, leading to the need for disambiguation of hashtags. We apply `TweetGrep` to sift through the mass of tweets grouped under a common hashtag, and identify the tweets that talk about the original sense of that hashtag. Similar to the task of retrieval of topical tweets, we learn this jointly with sentiment analysis. To the best of our knowledge, this is the first attempt to address the problem of hashtag disambiguation. We compare `TweetGrep` with two baselines, and it outperforms the more competitive baseline by average improvement of 6%.

The rest of the paper is organized as follows: We place the current work into perspective in Section 2. We formalise the retrieval and adaptation problems we study herein, detail the datasets we study, and set up the notations in Sections 3, 4 and 5 respectively. We describe our topic and sentiment baselines in Sections 6 and 7 respectively. Finally we develop our joint model in Section 8, and explain our experimental results in Section 9. We conclude our paper in Section 10 with possible future directions.

2 Related Work

2.1 Topic (Event) Detection

There exists a vast body of literature on *event* detection from Twitter. (Chierichetti et al. 2014) detect large-scale events impacting a significant number of users, e.g. a goal in the Football World Cup, by exploiting fluctuations in peer-to-peer and broadcast communication volumes. (Ritter et al. 2012) perform *open-domain* event detection, e.g. death of Steve Jobs, by inspecting trending named-entities in the tweets. (Ritter et al. 2015) enable an analyst to define an event *succinctly* with its past instances, and learn to classify the event-related tweets. (Tan et al. 2014) isolate trending *topics* by modeling departures from a background topic model. None of these works, however, exploit the sentiment polarity of events (topics), whenever available, to augment the detection.

2.2 Sentiment Analysis

(Xiang et al. 2014) demonstrate that supervised sentiment analysis of tweets (Mohammad et al. 2013) can further be improved by clustering the tweets into topical clusters, and performing sentiment analysis on each cluster independently. (Wang et al. 2011) treat individual hash-tags (#) as topics (events) and perform sentiment analysis, seeded with few tweet-level sentiment annotations. On the other extreme,(Go et al. 2009) learn to perform sentiment analysis by treating the *emoticons* present in tweets as labels. Extensions presented in (Barbosa et al. 2010) and (Liu et al. 2012) utilise additional human annotations, whenever available. These works rely on annotations at the granularity of individual tweets, and do not explore possibility of learning from topic level annotations.

2.3 Domain Adaptation for Sentiment Analysis

(Blitzer et al. 2007)'s work is an extension of Structural Correspondence Learning (first introduced in (Blitzer et al 2006), to sentiment analysis where pivot features are used to link the source and target domains. Further, the correlations between the pivot features and all other features are obtained by training linear pivot predictors to predict occurrences of each pivot in the unlabeled data from both domains. They evaluate their approach on a corpus of reviews for four different types of products from Amazon. Somewhat similar to this is the approach of (Tan et al. 2009), who pick out generalizable features that occur frequently in both domains and have similar occurring probability. (Glorot et al. 2011) beat the then state-of-the-art on the same dataset of Amazon product reviews from four domains, as in (Blitzer et al. 2007), using stacked de-noising auto-encoders. Their system is trained on labeled reviews from one source domain, and it requires unlabeled data from all the other domains to learn a meaningful representation for each review using an unsupervised fashion. In our setting, we intend to be able to adapt to any topic and we do not have access to all possible topics before hand.

2.4 Topic (Event) & Sentiment *Jointly*

There had been attempts at modeling topics (events/sub-events) and the associated sentiments *jointly*. (Hu et al. 2013) *segment* event-related tweet corpora into *sub-events* (e.g. proceedings in US presidential debate) and perform sentiment analysis - provided with sentiment annotations and segment alignments at the granularity of tweets. (Lin et al. 2012) jointly model the dynamics of topics and the associated sentiments, where the topics are essentially the dominant themes in the corpora. (Jiang et al. 2011) perform *target-dependent* sentiment analysis in a supervised setting, where the target is specified with a query (e.g. `"Windows 7"`). Our work further explores the theme of *weak-supervision*: alleviating the need for granular annotations by modeling the topics and sentiments *jointly*.

2.5 Hashtag Classification and Recommendation

(Kotsakos et al.) classify a single hashtag as a meme or an event. (Godin et al.) on the other hand, recommend a hashtag to a tweet to categorize it and hence enable easy search of tweets. While both these works are related to our task of hashtag disambiguation, it is very different from our work, because our goal is to retrieve the tweets that talk about the original sense of the hashtag from a stream of tweets that *all contain that hashtag*.

3 Premise

As mentioned in Section 1, the social media analyst specifies the topic-of-interest with a pair of queries, $\mathcal{Q}$ and $\mathcal{Q}^+$. Of these, $\mathcal{Q}$ retrieves *all* the topical tweets, *including false*

Table 1: Notations used in this work.

Notation	Interpretation
$\mathcal{T}_E^+$	Positive bag of tweets for topic E.
$\mathcal{T}_E \setminus \mathcal{T}_E^+$	Mixed bag of tweets for topic E.
$\mathcal{Q}$	Twitter Search queries used to retrieve $\mathcal{T}$
$\mathcal{Q}^+$	Twitter Search queries used to retrieve $\mathcal{T}^+$
$y^E(\tau)$	Topic labels $\in \{\pm 1\}$
$y^S(\tau)$	Sentiment labels $\in \{\pm 1\}$
$x^E(\tau)$	Feature representations for Topic
$x^S(\tau)$	Feature representations for Sentiment
w^E	Parameters for Topic
w^S	Parameters for Sentiment
$\tilde{p}_E$	Expected proportion of topical tweets
$\tilde{p}_{S,E}$	Expected proportion of topical tweets with positive polarity
$\tilde{p}_{S,\neg E}$	Expected proportion of non-topical tweets with positive polarity

positives. We denote the set of tweets retrieved with $\mathcal{Q}$ as $\mathcal{T}$. On the other hand, $\mathcal{Q}^+$ retrieves a *subset* of topical tweets, $\mathcal{T}^+ \subseteq \mathcal{T}$, and is guaranteed not to contain false positives. The goal of the learning process is to weed out false positives contained in the mixed bag, $\mathcal{T} \setminus \mathcal{T}^+$.

In practice, $\mathcal{Q}^+$ is often a specialisation of $\mathcal{Q}$: appending new clauses to $\mathcal{Q}$. For example, for CIVILWAR, the query $\mathcal{Q}$ is set to: `civil war`[3], and $\mathcal{Q}^+$ appends `captain america OR captainamerica` to $\mathcal{Q}$. One might wonder why we do not search using `civil war captain america OR captainamerica` to begin with - it is because all true positives might not necessarily contain explicit reference to `captain america`, but instead might talk about other characters or aspects of the movie. Hence our first query must be one that has high recall, even if it contains some false positives. This scheme is expressive enough to express a wide variety of topics. In another example, if the analyst is interested in public perceptions around FLIPKART's newly launched IMAGESEARCH[4] feature, she would set $\mathcal{Q}$ to `flipkart image search`. However, that includes buzz around a related bloggers' meet too, so in order to remove that, she would append `feature` such that $\mathcal{Q}^+$ would become `flipkart image search feature`.

To regulate the learning process, we further request the analyst to furnish three *expected* quantities, often easily obtained through domain expertise: $\tilde{p}_E$ quantifying her expectation of the fraction of topical tweets in the mixed bag $\mathcal{T} \setminus \mathcal{T}^+$, $\tilde{p}_{S,E}$ expressing her expectation of the fraction of topical tweets carrying positive sentiment, and $\tilde{p}_{S,\neg E}$ for that of non-topical tweets with positive sentiment. All the system inputs and parameters are defined concisely in Table 1. The regularisation process will be detailed in the following sections, and through extensive experiments, we will conclude that the learning process is resistant to noises in estimation of these quantities.

4 Datasets

4.1 Retrival of Topical Posts and Sentiment Analysis

To study the efficacy of our method for the task of retrieval of topical posts, we experiment with two types of data:

1. Tweets related to a wide spectrum of events: ranging from new features launches and strategic decisions by Indian e-commerce giants, to Stock Market crash in China.
2. Reviews, synthetically adapted to our setting, from publicly available Semeval 2014 dataset.

In the following, we briefly define and describe each of the datasets that we use for this task.

Flipkart's Image Search Launch (IMAGESEARCH). In July 2015, India's largest e-commerce portal, FLIPKART, announced the launch of Image Search that enables users to search products by clicking their pictures. $\mathcal{T}_E$, retrieved with $\mathcal{Q} \triangleq$ `flipkart image search`, additionally contained tweets pertaining to a bloggers' meet organised around this launch. We set $\mathcal{Q}^+ \triangleq$ `flipkart image search feature` to weed them out.

$\mathcal{T}_E$ contains tweets like:

"*#FlipkartImageSearch indiblogger meet at HRCIndia today! Excited! #Bangalore bloggers, looking forward to it!*".

On the other hand, $\mathcal{T}_E^+$ only contains tweets like:

"*Shopping got even better @Flipkart. Now You Can – POINT. SHOOT. BUY. Introducing Image Search on #Flipkart*".

Civil War, the Marvel Motion Picture (CIVILWAR). Tweets related to the upcoming Marvel motion picture, *Captain America: Civil war*, slated to be released in 2016, are of interest in the CIVILWAR dataset. While we retrieve $\mathcal{T}_E$ with $\mathcal{Q} \triangleq$ `civil war`, tweets related to the tragic events unfolding in Syria and the Mediterranean match the query, too:

"*83%: that's how much territory #Assad's regime has lost control of since #Syria's civil war began http://ow.ly/Rl19i*"

Additionally, recent allusion to American Civil War during US presidential debate and documentary film-maker and historian Ken Burns' re-mastered film are also retrieved.

Of these, we focus on the Marvel motion picture by retrieving $\mathcal{T}_E^+$ with $\mathcal{Q}^+ \triangleq$ `civil war captain america OR civil war captainamerica`.

Stock Market Crash in China (CRASH). The Stock Market Crash in China began with the popping of the investment bubble on 12 June, 2015. One third of the value of A-shares on the Shanghai Stock Exchange was lost within a month of that event. By 9 July, the Shanghai stock market had fallen 30% over three weeks. We retrieve $\mathcal{T}_E$ with $\mathcal{Q} \triangleq$ `china crash`, to avoid missing tweets like the following:

"*Japanese researchers think #China's GDP could crash to minus 20% in the next 5 years. http://ept.ms/1OciIiB pic.twitter.com/ia2pFNKjvi*"

We focus on $\mathcal{T}_E^+$ with $\mathcal{Q}^+ \triangleq$ `china crash market`.

[3] The syntax follows https://goo.gl/4NfP2A

[4] http://goo.gl/D5f5EZ

SemEval 2014 Restaurants data (ABSA). The restaurant dataset[5] for the task of Aspect based Sentiment Analysis consists of reviews belonging to the following aspects - food, service, ambience, price and miscellaneous. Out of these, we consider the first four - food (Food), service (Service), ambience (Ambience) and price (Price). The reviews are labeled with aspects and their associated sentiments. Each review can belong to more than one aspect, and we have a sentiment label per aspect. We synthetically make this dataset compatible to our experimental settings - we use the labels only for evaluation purposes, and employ keyword search to create the positive bag $\mathcal{T}_E^+$ for each aspect. For each of the aspects, the unlabeled bag consists of all reviews except the corresponding $\mathcal{T}_E^+$. The keywords used for each aspect are shown in Table 2.

Table 2: Queries for each of the aspects of the ABSA Restaurants dataset.

Aspect	**Query** $\mathcal{Q}^+$
Service	service OR staff OR waiter OR rude OR polite
Ambience	ambience OR decor OR environment
Price	price OR cost OR expensive OR cheap OR money
Food	food OR menu OR delicious OR tasty

4.2 Hashtag Disambiguation and Sentiment Analysis

For this task, we collect tweets related to certain viral hashtags and disambiguate them so as to alleviate the problem of hashtag overloading - the use of the same hashtag for multiple and morphed topics. The dataset used for this work, and the associated queries have been described below.

Paris Terror Attacks 2015 (Porteouverte). On 13 November 2015, a series of coordinated terrorist attacks occurred in Paris and its northern suburb, killing 130 people and injuring several others. The hashtag #porteouverte ("open door") was used by Parisians to offer shelter to those afraid to travel home after the attacks. But many people started using this hashtag to discuss the attack, to condemn it, to show solidarity and to express hope in humanity. As a result, the actually helpful tweets got drowned in the midst. We are interested in retrieving only those tweets which could be of any tangible help to the people stranded because of the Paris terror attacks. We retrieve $\mathcal{T}_E$ with $\mathcal{Q} \triangleq$ `#porteouverte`. However, it contained tweets such as

"*Thoughts with Paris today, very sad. #prayforparis #jesuisparis #porteouverte.*"

So, we employ another specific query $\mathcal{Q}^+ \triangleq$ `#porteouverte shelter` to obtain $\mathcal{T}_E^+$ that contains only tweets such as

"*If need a shelter in the 18th district, follow and DM, there is a #porteouverte here #AttentatsParis.*"

Table 3: Statistics of the datasets: sizes of the bags, rarity of the topic in $\mathcal{T} \setminus \mathcal{T}^+$, and the sentiment polarity across datasets (see Table 1 for the notations).

Dataset	$\lvert\mathcal{T}^+\rvert$	$\lvert\mathcal{T} \setminus \mathcal{T}^+\rvert$	$\tilde{p}_E$	$\tilde{p}_{S,E}$	$\tilde{p}_{S,\neg E}$
ImageSearch	113	727	0.63	0.61	0.33
CivilWar	304	305	0.35	0.25	0.17
Crash	627	497	0.81	0.25	0.03
Porteouverte	2630	248	0.157	0.149	0.604
ChennaiFloods	240	216	0.375	0.324	0.551
Service	261	1534	0.16	0.07	0.62
Ambience	49	1746	0.129	0.087	0.61
Price	198	1597	0.042	0.022	0.64
Food	487	1308	0.445	0.3577	0.3241

Chennai Floods 2015 (ChennaiFloods). Resulting from the heavy rainfall of the annual northeast monsoon in November-December 2015, the floods particularly hit hard the city of Chennai, India, killing more than 400 and displacing over 1.8 Million people. Twitter became the primary platform of communication - all tweets made in connection with the volunteer activities were tagged with #ChennaiRainsHelp and #ChennaiMicro so as to make it easier to search for. However, a lot of noise was included with time. Our goal is to filter out the noise and extract only those tweets that are in tune with the original sense of the hashtag and serves its original purpose, i.e. to share useful information, to communicate effectively and mobilize help for people affected in the floods. We retrieve $\mathcal{T}_E$ with $\mathcal{Q} \triangleq$ `#chennaimicro` OR `#chennairainshelp`. As expected, noisy tweets were also retrieved like the following:

"*So people used smartphones smartly to do something that technology always wants us to do. #chennairains #chennairainshelp*"

We extract the high precision, low recall $\mathcal{T}_E^+$ using the query $\mathcal{Q}^+ \triangleq$ `#chennaimicro need OR #chennairainshelp need` such that we now retrieve tweets only of the kind –

"*Okay guys! Calling out to ppl with supply trucks under them. Please inbox me. very urgent need to be met in Virudhachalam. #ChennaiMicro*"

Human Annotation. For each of these datasets except ABSA dataset (already labeled and publicly available), the queries are issued to the Twitter Search Web Interface via a proxy that we developed (and the results scraped), to alleviate restrictions around accessing tweets older than a week via the Twitter Search API. We obtain topic and sentiment annotations for all the tweets in the mixed bag, $\mathcal{T} \setminus \mathcal{T}^+$, through crowd-sourcing. 21 annotators from diverse backgrounds participated in this activity. Each $\mathcal{T} \setminus \mathcal{T}^+$ was broken down into blocks of 100 each, and sent to one of the annotators randomly. The protocol guarantees that each bag is annotated by a *group* of annotators, reducing the chance of bias in labeling. Each annotator was asked to provide two labels for a given tweet - a *topic* label denoting whether the tweet belongs to the topic or not, and a *sentiment* label denoting whether the tweet is positive or negative. An incentive of 1.5¢ per tweet was provided to the annotators. Table 3 captures sizes of the bags, rarity of the topic in $\mathcal{T} \setminus \mathcal{T}^+$, and

[5] http://goo.gl/3AXnIX

the associated sentiment proportions.

5 Notations

We begin with setting up the notations. For each tweet $\tau \in \mathcal{T}$, let $y^E(\tau) \in \{\pm 1\}$ denote the topicality of the tweet τ, with $y^E(\tau) = +1, \forall \tau \in \mathcal{T}^+$. For tweets in $\mathcal{T} \setminus \mathcal{T}^+$, $y^E(\tau)$ is a random variable endowed with the distribution $\Pr\left\{y^E(\tau) \mid x^E(\tau); w^E\right\}$; where $x^E(\tau) \in \Re^d$ denote the feature representation of τ, and $w^E \in \Re^d$ denote the corresponding parameter vector.

Similarly, let $y^S(\tau) \in \{\pm 1\}$ denote the sentiment labels for the tweet τ. Note that we do not deal with neutral sentiments in this work, mainly because non-subjective tweets are seldom of any importance for mining of opinion regarding topics of interest. Further, we endow $y^S(\tau), \forall \tau \in \mathcal{T}$ with the probability distribution $\Pr\left\{y^S(\tau) \mid x^S(\tau); w^S\right\}$, where $x^S(\tau) \in \Re^k$ and $w^S(\tau) \in \Re^k$ are the feature and parameter vectors, respectively. Table 1 summarises the notations used in this work.

6 Learning to Retrieve Topical Tweets - `BaseTopic`

In this section, we begin elaborating our very competitive baseline towards the retrieval of topical tweets, which is greatly motivated by (Ritter et al. 2015). It can be considered as a state-of-the-art for this task. In Section 7, we detail our baseline for sentiment analysis. Subsequently in Section 8 we develop a *joint* model `TweetGrep` that adapts the baseline sentiment analyser to the topic, and intertwines these two learning tasks. The approach described in this section will be compared against `TweetGrep`.

6.1 Learning with $\mathcal{T}^+$

In this work, we restrict the topic classifier to the maximum entropy family of classifiers, where $\langle \cdot, \cdot \rangle$ denotes the inner product:

$$\Pr\left\{y^E(\tau) \mid x^E(\tau); w^E\right\} = \frac{1}{1 + \exp[-\langle w^E, x^E(\tau)\rangle]}$$

Further, we posit a Gaussian prior over the parameter vector, $w^E \sim \mathcal{N}(0, \frac{1}{\lambda^E}\mathbb{I})$, where $\mathbb{I}$ is an appropriate identity matrix, and λ^E is the corresponding precision hyper-parameter, to aid our parameter estimation from potentially insufficient data.

We want to maximise the sum of the data log-likelihood, and include an l_2 norm term for regularization :

$$\sum_{\tau \in \mathcal{T}^+} \ln \Pr\left\{y^E(\tau) \mid x^E(\tau); w^E\right\} - \lambda^E \|w^E\|_2^2 \quad (1)$$

where $\|\cdot\|_2$ is the l_2 norm in $\Re^d$.

6.2 Learning with $\mathcal{T} \setminus \mathcal{T}^+$

Following the Expectation Regularisation framework (Mann and McCallum 2007) and (Ritter et al. 2015), we ensure that, $\hat{p}_E$, the *estimate* of fraction of topical tweets in $\mathcal{T} \setminus \mathcal{T}^+$, matches the supplied target *expectation*, $\tilde{p}_E$, in a KL divergence sense. This enables the learning process to leverage the unlabeled data, $\mathcal{T} \setminus \mathcal{T}^+$, and adds the following to the objective function (Equation 1):

$$\varrho KL(\tilde{p}_E \parallel \hat{p}_E) = \varrho \left\{\tilde{p}_E \ln \frac{\tilde{p}_E}{\hat{p}_E} + (1 - \tilde{p}_E) \ln \frac{1 - \tilde{p}_E}{1 - \hat{p}_E}\right\} \quad (2)$$

where the hyper-parameter ϱ controls the strength of the regularisation. The estimate, $\hat{p}_E$, is obtained as follows, where $\mathbb{1}_{\{\cdot\}}$ is an indicator random variable:

$$\begin{aligned}\hat{p}_E &= \frac{1}{|\mathcal{T} \setminus \mathcal{T}^+|} \sum_{\tau \in \mathcal{T} \setminus \mathcal{T}^+} \mathbb{E}_{y^E(\tau)}\left[\mathbb{1}_{\{y^E(\tau) = +1\}}\right] \\ &= \frac{1}{|\mathcal{T} \setminus \mathcal{T}^+|} \sum_{\tau \in \mathcal{T} \setminus \mathcal{T}^+} \Pr\left\{y^E(\tau) = +1 \mid x^E(\tau); w^E\right\}\end{aligned}$$

The gradient, ∇_{w^E} readily follows from (Ritter et al. 2015) and is omitted for the sake of brevity.

6.3 Feature Extraction

In our implementation, the feature representation, $x^E(\tau)$, consists of common nouns and verbs in tweets encoded as uni-grams, and rest of the words represented with their POS (Part Of Speech) tags for better generalisation. We use Ark Tweet NLP tool (Owoputi et al. 2013) for tokenization of tweets and for extracting POS tags.

7 Baseline Sentiment Analysis - `BaseSenti`

Starting with an *off-the-shelf* Sentiment Analyser, we would adapt it to capture topic-specific connotations. The process will be detailed in Section 8.

In line with the spirit of weak supervision, we pick our baseline from (Go et al. 2009). The sentiment analyser therein learns from hundreds of millions of tweets that contain *emoticons*, treating the sentiment conveyed by the emoticon as their labels. In particular, we train the baseline sentiment analyser on SENTIMENT140, a data-set containing 1.6 Million tweets from an assortment of domains. We call this model `BaseSenti`. Logistic regression, belonging to the same maximum entropy class of classifiers as our topic classifier, is used to learn the hyper-plane, w^0, which will further be *adapted* to capture topic-specific connotations in `TweetGrep` in Section 8.

Feature Extraction In our implementation, the feature representation, $x^S(\tau)$, consists of the uni-gram of tokens present in the tweet τ. For the tokenisation of tweets, we use the Ark Tweet NLP tool (Owoputi et al. 2013). After tokenization, the user-names following the `@` and the `URL`s are replaced with special tokens. Furthermore, as a pre-processing step, elongated words are normalised by replacing letters that repeat 2 or more times in a run with only 2 occurrences (e.g. *soooo* becomes *soo*). Frequency-based pruning is also employed.

8 `TweetGrep`

In this section, we elaborately describe our proposed *joint* optimisation framework, called `TweetGrep`.

8.1 Adapting `BaseSenti` with $\tilde{p}_{S,E}$ and $\tilde{p}_{S,\neg E}$

As is customary (Attenberg et al. 2009), we adapt w^0 for topical tweets in an *additive* fashion:

$$\Pr\left\{y^S(\tau) \mid y^E(\tau), x^S(\tau); w^S, w^0\right\} \\ = \begin{cases} \sigma\left(y^S(\tau) \times \langle w^S + w^0, x^S(\tau)\rangle\right) & \text{for} \quad y^E(\tau) = +1 \\ \sigma\left(y^S(\tau) \times \langle w^0, x^S(\tau)\rangle\right) & \text{for} \quad y^E(\tau) = -1 \end{cases} \tag{3}$$

where $\sigma(x) = \frac{1}{1+\exp(-x)}$ is an exponential family function, similar to the topic classifier. Simply put, $(w^S + w^0)$ is used as the parameter for topical tweets, while we use the `BaseSenti` parameter w^0 for the non-topical tweets. The intuition is that, we want to adapt `BaseSenti` only for topical tweets, while we would fall back to the `BaseSenti` parameters w^0 for other tweets. However, given that $y^E(\tau)$ is not known a priori for tweets in $\mathcal{T} \setminus \mathcal{T}^+$, learning w^S is not straightforward. To this end, we resort to an alternating optimisation that utilises the best estimates of $\Pr\left\{y^E(\tau)\right\}, \forall\tau \in \mathcal{T} \setminus \mathcal{T}^+$, obtained thus far, to estimate w^S, and then, in the next step, exploits the best estimates for $\Pr\left\{y^S(\tau)\right\}, \forall\tau \in \mathcal{T} \setminus \mathcal{T}^+$ to further estimate w^E. The opinion polarities, $\tilde{p}_{S,E}$ and $\tilde{p}_{S,\neg E}$, act as a bridge between these two learning problems, and regulate the transfer of learning.

To regulate the learning process, furthermore, we place a suitable Gaussian prior $w^S \sim \mathcal{N}(0, \frac{1}{\lambda^S}\mathbb{I})$. Mathematically, w^S is the minimiser of:

$$\varsigma \times KL(\tilde{p}_{S,E} \parallel \hat{p}_{S,E}) + \vartheta \times KL(\tilde{p}_{S,\neg E} \parallel \hat{p}_{S,\neg E}) \\ + \lambda^S \|w^S\|_2^2 \tag{4}$$

where ς and ϑ are hyper-parameters controlling the strength of regularisation.

In order to minimise $\varsigma \times KL(\tilde{p}_{S,E} \parallel \hat{p}_{S,E}) + \vartheta \times KL(\tilde{p}_{S,\neg E} \parallel \hat{p}_{S,\neg E})$, we need to obtain the estimates $\hat{p}_{S,E}$ and $\hat{p}_{S,\neg E}$ (See Equation 3 for more clarity) which are as follows:

$$\hat{p}_{S,E} = \frac{1}{|\mathcal{T} \setminus \mathcal{T}^+|} \sum_{\tau \in \mathcal{T} \setminus \mathcal{T}^+} \Pr\left\{y^E(\tau) = +1\right\} \\ \times \Pr\left\{y^S(\tau) = +1 \mid y^E(\tau) = +1\right\} \\ = \frac{1}{|\mathcal{T} \setminus \mathcal{T}^+|} \sum_{\tau \in \mathcal{T} \setminus \mathcal{T}^+} \frac{1}{1 + \exp[-y^E(\tau) \times \langle w^E, x^E(\tau)\rangle]} \\ \times \frac{1}{1 + \exp[-y^S(\tau) \times \langle (w^S + w^0), x^S(\tau)\rangle]} \tag{5}$$

Similarly, estimate of the proportion of positive tweets which are non-topical:

$$\hat{p}_{S,\neg E} \\ = \frac{1}{|\mathcal{T} \setminus \mathcal{T}^+|} \sum_{\tau \in \mathcal{T} \setminus \mathcal{T}^+} (1 - \Pr\left\{y^E(\tau) = +1\right\}) \times \\ \Pr\left\{y^S(\tau) = +1 \mid y^E(\tau) = -1\right\} \\ = \frac{1}{|\mathcal{T} \setminus \mathcal{T}^+|} \sum_{\tau \in \mathcal{T} \setminus \mathcal{T}^+} \frac{\exp[-y^E(\tau) \times \langle w^E, x^E(\tau)\rangle]}{1 + \exp[-y^E(\tau) \times \langle w^E, x^E(\tau)\rangle]}) \times \\ \frac{1}{1 + \exp[-y^S(\tau) \times \langle w^0, x^S(\tau)\rangle]} \tag{6}$$

8.2 Learning to Retrieve and to Adapt *Jointly*

Combining the topic terms (Equations 1 and 2) and the sentiment terms (Equation 4) mentioned earlier, the joint objective function becomes–
Maximize:

$$\sum_{\tau \in \mathcal{T}^+} \ln \Pr\left\{y^E(\tau) \mid x^E(\tau); w^E\right\} - \varrho \times KL(\tilde{p}_E \parallel \hat{p}_E) \\ - \lambda^E \|w^E\|_2^2 - \lambda^S \|w^S\|_2^2 - \varsigma \times KL(\tilde{p}_{S,E} \parallel \hat{p}_{S|E}) \\ - \vartheta \times KL(\tilde{p}_{S,\neg E} \parallel \hat{p}_{S|\neg E}) \tag{7}$$

While the gradients for the topic terms are straightforward and follow from (Ritter et al. 2015), for the sake of completeness, we present the gradients $\nabla_{w^E} KL(\tilde{p}_{S,E} \parallel \hat{p}_{S,E})$ and $\nabla_{w^S} KL(\tilde{p}_{S,\neg E} \parallel \hat{p}_{S,\neg E})$. We skip the derivation due to lack of space. The final forms of the gradients are as follows:

$$\nabla_{w^E} KL(\tilde{p}_{S,E} \parallel \hat{p}_{S,E}) \\ = \frac{1}{|\mathcal{T} \setminus \mathcal{T}^+|} \left(\frac{1 - \tilde{p}_{S,E}}{1 - \hat{p}_{S,E}} - \frac{\tilde{p}_{S,E}}{\hat{p}_{S,E}}\right) \times \\ \sum_{\tau \in \mathcal{T} \setminus \mathcal{T}^+} \Pr\left\{y^E(\tau) = +1\right\} \times (1 - \Pr\left\{y^E(\tau) = +1\right\}) \times \\ \Pr\left\{y^S(\tau) = +1 \mid y^E(\tau) = +1\right\} \times x^E(\tau) \tag{8}$$

$$\nabla_{w^S} KL(\tilde{p}_{S,\neg E} \parallel \hat{p}_{S,\neg E}) \\ = \frac{1}{|\mathcal{T} \setminus \mathcal{T}^+|} \left(\frac{1 - \tilde{p}_{S,E}}{1 - \hat{p}_{S,E}} - \frac{\tilde{p}_{S,E}}{\hat{p}_{S,E}}\right) \times \\ \sum_{\tau \in \mathcal{T} \setminus \mathcal{T}^+} \Pr\left\{y^S(\tau) = +1 \mid x^S(\tau), y^E(\tau) = +1, w^0, w^S\right\} \times \\ (1 - \Pr\left\{y^S(\tau) = +1 \mid x^S(\tau), y^E(\tau) = +1, w^0, w^S\right\}) \times \\ \Pr\left\{y^E(\tau) = +1 \mid x^E(\tau); w^E\right\} \times x^S(\tau) \tag{9}$$

Armed with these gradients (Equations 8 and 9), the joint optimisations are carried out in an alternating fashion using L-BFGS (Byrd et al. 1995) until convergence. Random restarts are employed to scout for better optima. In practice, the hyper-parameters, $\{\varrho, \varsigma, \vartheta\}$ are optimized on a held-out validation set using Bayesian hyper-parameter tuning tool Spearmint (Snoek et al. 2012). λ^E and λ^S are simply set to 100 following (Ritter et al. 2015).

9 Evaluation

In this section, we elaborate the performance of `TweetGrep` on two applications: retrieval of topical tweet/posts and hashtag disambiguation, along with sentiment analysis for both.

9.1 Retrieval of Topical Posts and Sentiment Analysis

The first application aims at retrieving the relevant topical posts from the unlabeled bag of tweets. Here `TweetGrep` is compared with `BaseTopic` as described in Section 6. We create the gold-standard annotations as mentioned earlier in Section 4. The performances of the models are compared with the human annotations in terms of true and false positive rates (TPR and FPR, respectively) and the area under the curve (AUC) is reported.

Figure 1 (upper panel) shows the ROC curves of two competing models for different datasets. The corresponding values for AUC are reported in Table 4 (second and third columns).

We observe in Table 4 that with respect to the `BaseTopic`, the improvement of `TweetGrep` is maximum for CIVILWAR (14.29%), which is followed by IMAGESEARCH (5.34%), AMBIENCE (5.32%), PRICE (4.71%), CRASH (3.11%), FOOD (1.04%) and SERVICE (1.01%). The average improvement of `TweetGrep` is 4.97% with respect to `BaseTopic` irrespective of the datasets. The reason behind the best performance on the CIVILWAR dataset may be as follows: the polarity distribution is such that almost all positive tweets are topical and all negative tweets are non-topical, thereby leading to sentiment greatly helping the task of topical retrieval.

Similarly, for sentiment analysis, we compare `TweetGrep` with baseline model `BaseSenti` (described in Section 7) in terms of TPR and FPR. The lower panel of Figure 1 shows the ROC curves of two models for different datasets, and the AUC values are reported in Table 4 (forth and fifth columns).

From Table 4, we note that the maximum improvement of `TweetGrep` compared to `BaseSenti` occurs for CRASH (9.68%), followed by IMAGESEARCH (8.55%), CIVILWAR (7.20%), FOOD (6.24%), PRICE (6.15%), AMBIENCE (5.62%) and SERVICE (4.94%). The average improvement of `TweetGrep` is significantly higher (6.91%) than `BaseSenti` irrespective of the datasets. From the tweets dataset, CIVILWAR performs slightly worse than the others, which can be attributed to the fact that there is a lot of variety among the non-tropical tweets (civil war in Syria, documentary film, American civil war, etc. - see Section 4), each with their own associated aggregate sentiments, thereby lessening the scope of salvaging the joint learning by the sentiment classifier. In the ABSA dataset, the performance of `TweetGrep` is comparatively poor for SERVICE, although the improvement is significant nonetheless.

Examples of high-confidence extractions are presented in Tables 6 and 7 respectively for the two sub-tasks - retrieval of topical tweets and sentiment analysis - challenging samples that are misclassified by the baselines `BaseTopic` and `BaseSenti` but correctly identified by `TweetGrep`.

Robustness In Figure 2, we vary each of the parameters of the model while keeping the others fixed and shows the plot for AUC. The true priors for CIVILWAR dataset are $\tilde{p}_E = 0.3$, $\tilde{p}_{S,E} = 0.25$ and $\tilde{p}_{S,\neg E} = 0.17$. Figure 2(a) is a plot of AUC for CIVILWAR dataset, by varying $\tilde{p}_E$, keeping $\tilde{p}_{S,E}$ and $\tilde{p}_{S,\neg E}$ fixed at their true values. Similarly, Figure 2(b) shows plot of AUC by varying $\tilde{p}_{S,E}$, setting $\tilde{p}_E$ and $\tilde{p}_{S,\neg E}$ fixed, while Figure 2(c) varies $\tilde{p}_{S,\neg E}$ with $\tilde{p}_E$ and $\tilde{p}_{S,E}$ fixed. The variations range between ± 0.05 of the true prior, with an interval of 0.01. As is evident from the figures, the performance of `TweetGrep` remains almost uniform throughout the range of deviation. The results for the other datasets have similar patterns and have been omitted for the sake of brevity. This demonstrates that the performance of our system is robust, despite having a number of parameters, and works considerably well even when the true priors are not accurately known.

9.2 Hashtag Disambiguation and Sentiment Analysis

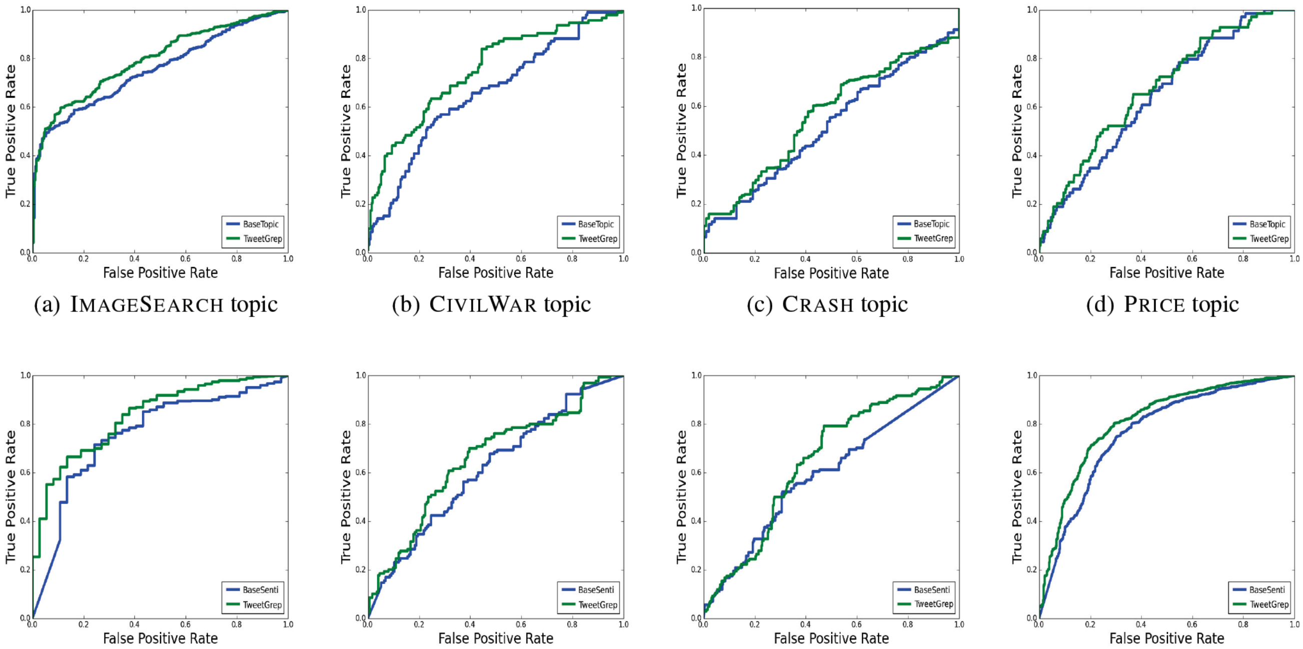
Hashtag disambiguation is the task of retrieval of tweets pertaining to the original sense of a particular hashtag, from a huge stream of tagged tweets, thereby solving the problem of overloading of a hashtag. To the best of our knowledge, this task is addressed for the first time in this paper. Therefore, for the purpose of comparison, we design two baseline models and show that `TweetGrep` performs significantly better than these baselines.

`BaseTopic` (introduced in Section 6) is applied to hashtag disambiguation, under similar settings as described for the task of retrieval of topical tweets. We collect $\mathcal{T}$ where $\mathcal{Q}$ consists of the hashtag itself. In order to obtain $\mathcal{T}^+$, we use an additional set of keyword(s) $\mathcal{Q}^+$. We are tasked with retrieving the tweets that belong to the original or desired sense of the hashtag in question.

`BaseHashtag` is another newly introduced baseline, taking inspiration from (Olteanu et al. 2014) and (Magdy et al. 2014). Similar to `BaseTopic` that is keyword-based; here too, we start with the high-precision, low-recall yielding keywords for retrieving the surely positive tweets, and then expand the set of keywords by adding their Wordnet synonyms. WordNet (Miller 1995) is a huge lexical database that can capture semantic information. The Wordnet-expanded keyword set is thereby used to filter out topical tweets from the unlabeled tweet bag $\mathcal{T} \setminus \mathcal{T}^+$. We use the Python package NLTK (Bird et al. 2009) to find Wordnet synonyms.

For the task of Sentiment Analysis, `TweetGrep` is compared with `BaseSenti` as before.

For this hashtag disambiguation task, we compare the results of `TweetGrep` with `BaseTopic` and `BaseHashtag`. Table 5 compares the three methods for



(a) ImageSearch topic (b) CivilWar topic (c) Crash topic (d) Price topic

(e) ImageSearch sentiment (f) CivilWar sentiment (g) Crash sentiment (h) Price sentiment

Figure 1: ROC for retrieval of topical tweets (upper panel) and their sentiment analysis (lower panel). For retrieval of topical tweets, `TweetGrep` is compared with `BaseTopic`, while for sentiment analysis, it is compared with `BaseSenti`. From the ABSA dataset, we show the ROC of only one aspect Price for the sake of brevity. The other aspects show similar characteristics, and their AUC values have been reported in Table 4.

Table 4: Comparison of `TweetGrep` with baselines with respect to their AUC values in the two tasks - Retrieval of topical tweets and Sentiment Analysis.

Dataset	**Topic**		**Sentiment**	
	`BaseTopic`	`TweetGrep`	`BaseSenti`	`TweetGrep`
ImageSearch	0.7537	0.7940	0.7575	0.8223
CivilWar	0.6548	0.7484	0.6136	0.6578
Crash	0.5677	0.5854	0.5909	0.6481
Service	0.4250	0.4675	0.7591	0.7966
Ambience	0.4643	0.4890	0.7668	0.8099
Price	0.6451	0.6755	0.7651	0.8122
Food	0.5552	0.5610	0.7542	0.8013

Table 5: AUC values of the competing methods in the task of hashtag disambiguation.

Dataset	`BaseTopic`	`BaseHashtag`	`TweetGrep`
PorteOuverte	0.7866	0.5128	0.8493
ChennaiFloods	0.6921	0.5197	0.7202

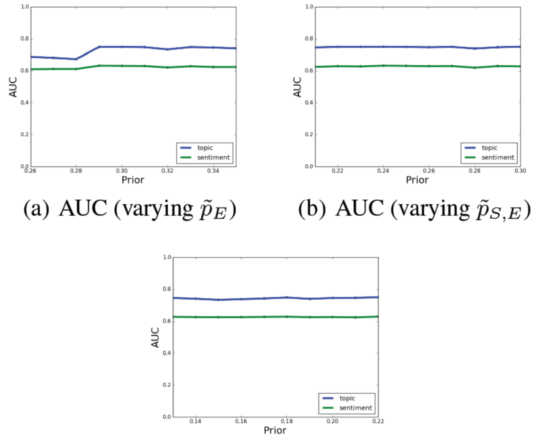
this task. Figure 3 shows the ROC curve, contrasting `TweetGrep` with the `BaseTopic` (upper panel). Please note that ROC curve of `BaseHashtag` is not shown because we do not deal with probability outcomes in this method, and hence ROC curve does not make much sense. As we can observe, `TweetGrep` beats `BaseHashtag` by a huge margin for both PorteOuverte and ChennaiFloods. This shows that seed expansion without context does not yield satisfactory results. The Wordnet expanded query set for PorteOuverte consists of {shelter, tax shelter, protection}, while that for ChennaiFloods contains {need, motivation, necessitate, want, indigence}. `BaseTopic` presents a strong baseline for our system by performing competitively. While the performance of `TweetGrep` is 7.97% better than `BaseTopic` on PorteOuverte, and that on ChennaiFloods is 4.06%. The poorer performance of `TweetGrep` on ChennaiFloods in comparison to PorteOuverte could be because of the following: the size of the training set $\mathcal{T}_E^+$ for PorteOuverte is 2630, while that for ChennaiFloods is a mere 240. Hence the learning was naturally better in the former. Secondly, we observe that the

Table 6: Anecdotal examples for topical retrieval task

Topic	Tweet	`BaseTopic`	`TweetGrep`	Gold
Crash	2 people killed in helicopter crash in mountain of SW China's Guangxi onmorning of Sept 21	Relevant	Not Relevant	Not Relevant
ImageSearch	All you gotta do to shop now is, Shoot! A picture!! #FlipkartImageSearch Indimeet	Not Relevant	Relevant	Relevant
Service	the cream cheeses are out of this world and i love that coffee!!	Relevant	Not Relevant	Not Relevant
PorteOuverte	If you're in the XII arrondissement and looking for a safe place to stay we can welcome you. #PorteOuverte"	Not Relevant	Relevant	Relevant

Table 7: Anecdotal examples for sentiment analysis task

Topic	Tweet	`BaseSenti`	`TweetGrep`	Gold
Crash	Japanese Researchers Think China's GDP Could Crash to Minus 20 Percent in Next 5 Years	Positive	Negative	Negative
ImageSearch	Latest from PlanetRetail: FLIPKART steals a march with in-app image search http://ift.tt/1M7HFY9 #Retail	Negative	Positive	Positive
Service	To my right, the hostess stood over a busboy and hissed rapido, rapido as he tried to clear and re-set a table for six.	Positive	Negative	Negative
PorteOuverte	The first of the storm. Translation of Isis claim of Paris attacks https://t.co/siDCgVDMxv #ParisAttacks #PorteOuverte	Positive	Negative	Negative



(a) AUC (varying $\tilde{p}_E$)

(b) AUC (varying $\tilde{p}_{S,E}$)

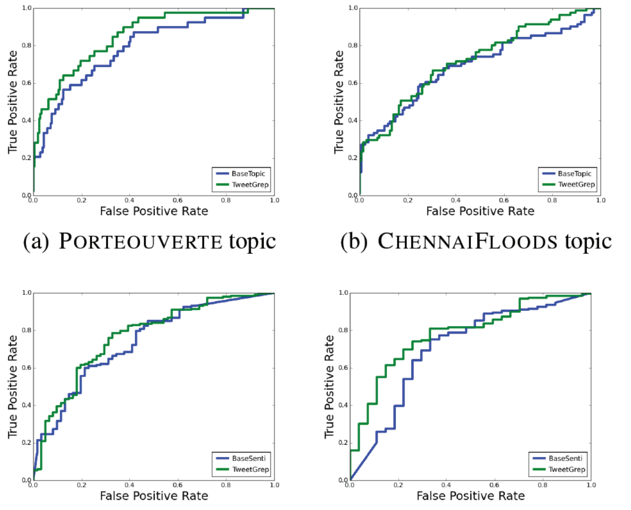
(c) AUC (varying $\tilde{p}_{S,\neg E}$)

Figure 2: Robustness of `TweetGrep` with respect to priors for the dataset CivilWar. True $\tilde{p}_E = 0.3$, $\tilde{p}_{S,E} = 0.25$, $\tilde{p}_{S,\neg E} = 0.17$.

larger the difference between $\tilde{p}_{S,E}$ and $\tilde{p}_{S,\neg E}$, the better is the influence of sentiment on hashtag disambiguation task. As we can see from Table 3, $\tilde{p}_{S,E}$ and $\tilde{p}_{S,\neg E}$ for PorteOuverte are 0.149 and 0.604 respectively, while that of ChennaiFloods are 0.324 and 0.551 respectively.

`BaseSenti` achieves an AUC of 0.7384 while `TweetGrep` gets a 3% improvement as 0.7667 on PorteOuverte dataset. For the ChennaiFloods dataset, the sentiment baseline and `TweetGrep` achieves 0.7054 and 0.7797 respectively. Figures 3(c) and 3(d) show the ROC for sentiment analysis and can be referred to, for finer details of the performance of the competing models.

We further observe for both topical posts retrieval and hashtag disambiguation, the improvement in performance of `TweetGrep` in constrast to the baselines is complementary for topic retrieval/hashtag disambiguation and sentiment analysis. For example, `TweetGrep` performs better for hashtag disambiguation for PorteOuverte while for ChennaiFloods the sentiment performance is better. Similarly, for the retrieval of topical posts, `TweetGrep` performs the best for CivilWar while its performance for sentiment is the worst of CivilWar among the tweets dataset.



(a) PorteOuverte topic

(b) ChennaiFloods topic

(c) PorteOuverte sentiment

(d) ChennaiFloods sentiment

Figure 3: ROC for hashtag disambiguation. The upper panel (topic retrieval) shows comparison of `TweetGrep` and `BaseTopic`. The lower panel (sentiment analysis) shows its performance with respect to `BaseSenti`.

10 Conclusion

To effectively gauge public opinion around a plethora of topics as portrayed in Social Media, the analyst would have to be equipped with quick and easy ways of training topic *and* sentiment classifiers. `TweetGrep`, the proposed *joint, weakly-supervised* model detailed herein, significantly outperformed state-of-the-art individual models in an array of experiments with user-generated content. We also applied `TweetGrep` to the hitherto underexplored task of hashtag disambiguation and demonstrated its efficacy.

11 Acknowledgement

We would like to thank Amod Malviya, the erstwhile Chief Technology Officer of Flipkart, for seeding and fostering the work.

References

Attenberg, J.; Weinberger, K.; Dasgupta, A.; Smola, A.; and Zinkevich, M. 2009. Collaborative email-spam filtering with the hashing-trick. In *CEAS, California, USA*, 1–4.

Barbosa, L., and Feng, J. 2010. Robust sentiment detection on twitter from biased and noisy data. In *COLING, Beijing, China*, 36–44.

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition.

Blitzer, J.; Dredze, M.; and Pereira, F. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL, Prague*, 187–205.

Blitzer, J.; McDonald, R.; and Pereira, F. 2006. Domain adaptation with structural correspondence learning. In *EMNLP, Sydney, Australia*, 120–128.

Byrd, R. H.; Lu, P.; Nocedal, J.; and Zhu, C. 1995. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.* 16:1190–1208.

Chierichetti, F.; Kleinberg, J. M.; Kumar, R.; Mahdian, M.; and Pandey, S. 2014. Event detection via communication pattern analysis. In *ICWSM, Oxford, U.K.*, 51–60.

Glorot, X.; Bordes, A.; and Bengio, Y. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML, Washington, USA*, 513–520.

Go, A.; Bhayani, R.; and Huang, L. 2009. Twitter sentiment classification using distant supervision. In *Final Projects from CS224N for Spring 2008/2009 at The Stanford NLP Group*, 1–6.

Godin, F.; Slavkovikj, V.; De Neve, W.; Schrauwen, B.; and Van de Walle, R. 2013. Using topic models for twitter hashtag recommendation. In *WWW*, 593–596.

He, Y.; Lin, C.; Gao, W.; and Wong, K. 2012. Tracking sentiment and topic dynamics from social media. In *ICWSM, Dublin, Ireland*, 483–486.

Hu, Y.; Wang, F.; and Kambhampati, S. 2013. Listening to the crowd: automated analysis of events via aggregated twitter sentiment. In *IJCAI, Beijing, China*, 2640–2646.

Jiang, L.; Yu, M.; Zhou, M.; Liu, X.; and Zhao, T. 2011. Target-dependent twitter sentiment classification. In *ACL, Portland, USA*, 151–160.

Kotsakos, D.; Sakkos, P.; Katakis, I.; and Gunopulos, D. 2014. #tag: Meme or event? In *ASONAM, Beijing, China*, 391–394.

Liu, K.-L.; Li, W.-J.; and Guo, M. 2012. Emoticon smoothed language models for twitter sentiment analysis. In *AAAI, Ontario, Canada*, 1678–1684.

Magdy, W., and Elsayed, T. 2014. Adaptive method for following dynamic topics on twitter. In *ICWSM, Michigan, USA*.

Mann, G. S., and McCallum, A. 2007. Simple, robust, scalable semi-supervised learning via expectation regularization. In *ICML, Corvallis, USA*, 593–600.

Miller, G. A. 1995. Wordnet: A lexical database for english. *Commun. ACM* 38:39–41.

Mohammad, S.; Kiritchenko, S.; and Zhu, X. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *SemEval, Atlanta, USA*, 321–327.

Olteanu, A.; Castillo, C.; Diaz, F.; and Vieweg, S. 2014. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *ICWSM, Michigan, USA*.

Owoputi, O.; OConnor, B.; Dyer, C.; Gimpel, K.; Schneider, N.; and Smith, N. A. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *NAACL-HLT, Atlanta, USA*, 380–390.

Pontiki, M.; Galanis, D.; Pavlopoulos, J.; Papageorgiou, H.; Androutsopoulos, I.; and Manandhar, S. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *SemEval, Dublin, Ireland*, 27–35.

Ritter, A.; Etzioni, O.; Clark, S.; et al. 2012. Open domain event extraction from twitter. In *ACM SIGKDD, Beijing, China*, 1104–1112.

Ritter, A.; Wright, E.; Casey, W.; and Mitchell, T. 2015. Weakly supervised extraction of computer security events from twitter. In *WWW, Florence, Italy*, 896–905.

Snoek, J.; Larochelle, H.; and Adams, R. P. 2012. Practical bayesian optimization of machine learning algorithms. In *NIPS, Nevada, USA*. 2951–2959.

Tan, S.; Cheng, X.; Wang, Y.; and Xu, H. 2009. Adapting naive bayes to domain adaptation for sentiment analysis. In *ECIR, Toulouse, France*, 337–349.

Tan, S.; Li, Y.; Sun, H.; Guan, Z.; Yan, X.; Bu, J.; Chen, C.; and He, X. 2014. Interpreting the public sentiment variations on twitter. *IEEE TKDE* 6:1158–1170.

Wang, X.; Wei, F.; Liu, X.; Zhou, M.; and Zhang, M. 2011. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *CIKM, Scotland, UK*, 1031–1040.

Xiang, B., and Zhou, L. 2014. Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training. In *ACL, Maryland, USA*, 434–439.